ENGINEERING Foods
https://doi.org/10.2738/ENGF.2026.0012

Review

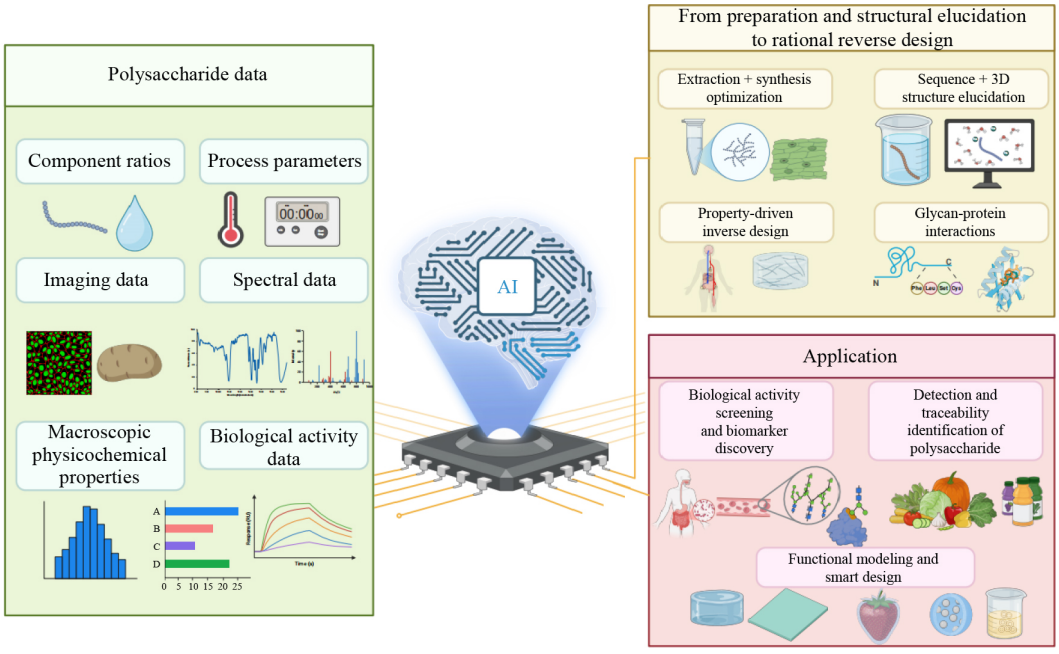
Next-Generation Glycan Research: Artificial Intelligence-Driven Extraction, Structural Elucidation, Food and Biomedical Applications

Jiachen Liang*, Abdul Mueed*, Lijun You (✉)

School of Food Science and Engineering, South China University of Technology, Guangzhou 510640, China

© The Author(s) 2026. This article is published by Higher Education Press.

GRAPHICAL ABSTRACT



HIGHLIGHTS

- AI integrates multimodal data for efficient glycan analysis.
- Machine learning optimizes polysaccharide extraction and synthesis.
- AI predicts glycan sequences, conformations, and molecular interactions.
- Data-driven models accelerate biomarker discovery and functional screening.
- AI supports rational design of polysaccharide-based food and biomaterials.

ABSTRACT Glycans are structurally complex and functionally diverse biomacromolecules that underpin key biological processes (including energy metabolism, cellular communication, and disease development) and food applications (including hydrogels, films, coatings and emulsions). Unlike nucleic acids and proteins, glycans are synthesized in a non-template-driven manner, resulting in extensive branching, isomerism, and heterogeneity that complicate their structural characterization and functional interpretation. These challenges are further compounded by the limited capacity of conventional experimental approaches to handle high-dimensional, nonlinear, and complex datasets. Recent advances in artificial intelligence (AI), particularly machine learning and deep learning are transforming this landscape. In this review,

Received 21 May 2026; accepted 7 Oct 2026

E-mail: feyoulijun@scut.edu.cn (You L.).

* These authors contributed equally to this work.

we examine emerging data resources, algorithmic frameworks, and benchmarking strategies that support AI-driven glycan research. We highlight how these approaches enable end-to-end integration across the glycan research pipeline, from experimental design and structural elucidation to predictive modeling and rational design. AI methods have demonstrated strong performance in decoding glycan sequences from spectroscopic and nanopore data, predicting three-dimensional conformations, and modeling glycan-protein interactions. Furthermore, AI enables inverse design strategies based on physicochemical and functional constraints. Beyond structural biology and food application, AI is accelerating biomarker discovery, functional screening, and the development of polysaccharide-based materials. Together, these advances signal a shift toward predictive and precision glycol-science, positioning AI as a central driver of innovation in this field.

KEYWORDS Machine learning; Sequence; Glycan; Polysaccharide; Application

1 Introduction

According to International Union of Pure and Applied Chemistry (IUPAC) nomenclature, polysaccharides (glycans) are macromolecules consisting of a large number of monosaccharide residues linked by glycosidic bonds, with the term "polysaccharide" commonly applied to chains containing more than ten monosaccharide residues. However, "glycan" is also used for the saccharide component of glycoproteins regardless of chain length. Therefore, both terms are retained in this review: "polysaccharide" primarily refers to independent carbohydrate macromolecules, whereas "glycan" is used for glycoconjugate-associated carbohydrate chains and glycan-specific datasets or studies. These structurally diverse carbohydrates participate in a wide range of biological processes, including cell adhesion, immune regulation, molecular recognition, host–microbe interactions, and intracellular signaling [1,2]. Beyond their essential biological functions, glycans possess remarkable physicochemical characteristics, such as biocompatibility, biodegradability, gel-forming capacity, rheological tunability, and molecular self-assembly behavior, which have led to their widespread applications in food systems, biomedical engineering, pharmaceutical delivery, tissue scaffolding, intelligent packaging, and advanced biomaterials [3,4]. In food science, polysaccharides are extensively utilized as stabilizers, emulsifiers, thickeners, edible coatings, and fat replacers, while numerous natural glycans exhibit antioxidant, immunomodulatory, hypoglycemic, hypolipidemic and prebiotic activities, highlighting their substantial potential for functional foods and precision nutrition [4,5].

Despite their significances, glycoscience has progressed more slowly than genomics and proteomics because glycans are synthesized through non-template driven pathways, resulting in extensive branching, linkage variability, stereochemical diversity, and structural heterogeneity [1,6]. These characteristics make glycan structural elucidation and structure-function analysis extremely challenging. Conventional analytical techniques, including nuclear magnetic resonance (NMR), mass spectrometry (MS), chromatography, and crystallographic approaches, provide valuable structural information but remain limited in resolving highly complex glycan structure and dynamic conformations [1,2,7]. Furthermore, the rapidly increasing volume and complexity of glycomics data have exceeded the capabilities of traditional computational and statistical methods.

Recent advances in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), are transforming glycoscience by enabling efficient analysis of high-dimensional and heterogeneous datasets (Fig. 1). AI particularly ML and DL are rapidly transforming in food science and biomacromolecular research. AI-driven approaches have demonstrated remarkable capability in handling high-dimensional, nonlinear and heterogeneous datasets generated from spectroscopy, imaging, omics and molecular characterization technologies. In recent years, ML algorithms have been increasingly applied in food processing, quality evaluation, biomarker discovery, structural elucidation and intelligent material design. For example, AI-driven approaches have demonstrated strong potential in glycan extraction optimization, sequence elucidation, structural prediction, glycan-protein interaction analysis, biomarker discovery and intelligent design of glycan-based materials [8–11]. In particular, deep learning models combined with MS, NMR nanopore sensing, and glycomics technologies have significantly improved the accuracy and throughput of glycan characterization and functional prediction [10,11]. However, compared with protein and other biomacromolecules, AI-assisted glycan research remains relatively underdeveloped and insufficiently summarized.

As shown in Fig. 1A, current AI-related research in food science is mainly concentrated in areas such as classification, food quality, computer vision, fermentation, and chemometrics, demonstrating the broad integration of AI technologies into food and biomolecular science. Fig. 1B further illustrates the exponential increase in AI and ML related publications during the past decade, reflecting the rapidly growing scientific and industrial interest in intelligent data-driven technologies. Nevertheless, Fig. 1C indicates that existing review studies remain largely focused on proteins, food quality detection and general food engineering applications

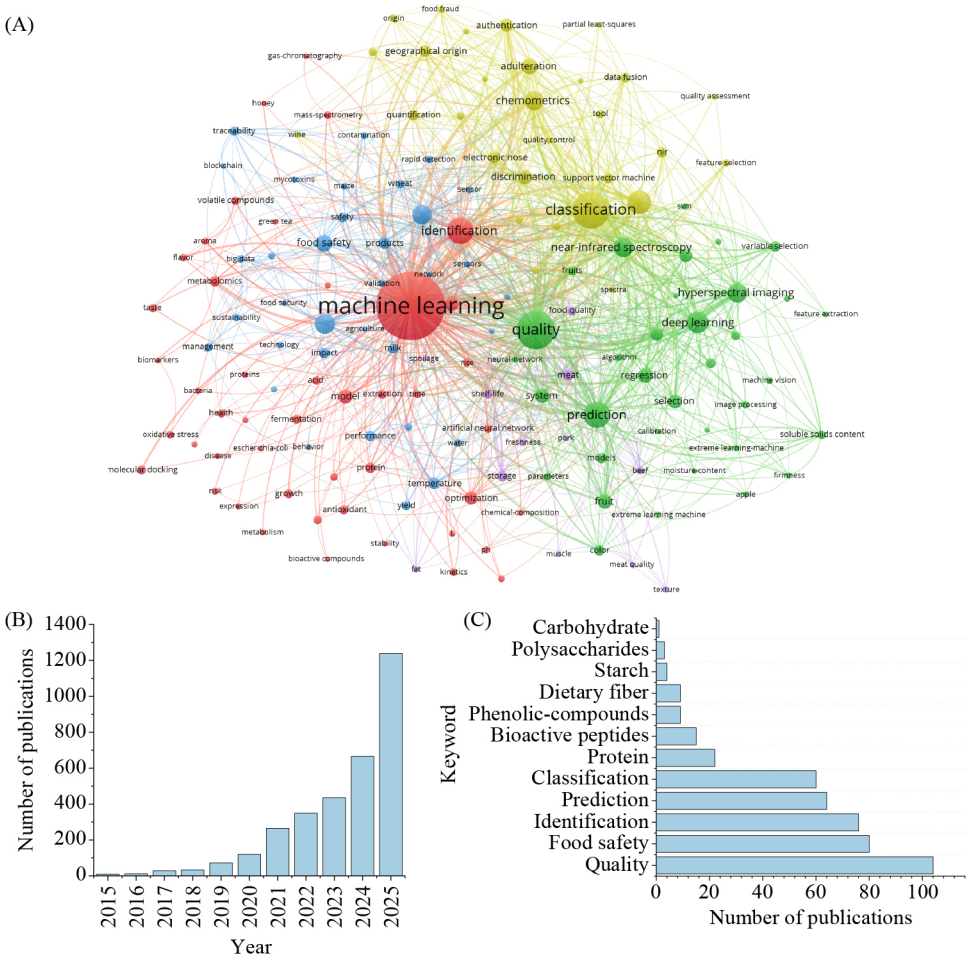


Fig. 1 Research progress of AI in the field of food science and engineering. (A) Hotspot map of ML-related keywords (produced by VOSviewer). (B) Number of literatures related to ML and AI from 2015 to 2025. (C) Number of review literature with different keywords.

Note: the literature search was conducted using the Web of Science Core Collection, specifically the Science Citation Index Expanded (SCI-EXPANDED) and Social Sciences Citation Index (SSCI). The search terms were "artificial intelligence" OR "machine learning". Only articles and reviews published between 2015 and 2025 in the field of food science and engineering were included.

whereas comprehensive reviews dedicated to AI-assisted glycan research are still limited. This research gap highlights the urgent need for a systematic overview of AI-driven strategies in glycoscience, particularly in the areas of glycan extraction, structural elucidation, molecular interaction prediction and functional material design. Therefore, this review provides a comprehensive overview of recent advances in AI-assisted glycan research. Particular emphasis is placed on AI-driven strategies for glycan extraction, structural elucidation, three-dimensional (3D) structure prediction, glycan-protein interaction analysis, biomarker discovery and functional material design. In addition, current challenges and future perspectives for integrating AI with glycoscience are discussed to promote the development of predictive, precision and mechanism-guided glycan research in food science, biomedicine, and advanced materials.

2 Challenges in the Polysaccharide Field

2.1 Limitations of molecular-level characterization methods

As one of the three major classes of biological macromolecules, glycans remain much less understood than proteins and nucleic acids in terms of chemical structure, sequence composition, 3D conformation, and intermolecular interactions. This gap mainly arises from their extraordinary structural complexity and heterogeneity, together with the limitations of current experimental and computational methods (Table 1 and Fig. 2). As a result, the slow progress in glycan structural elucidation has constrained the rational *de novo* design of glycans and hindered deeper understanding of their biological functions and applications.

Table 1 Experimental characterization and theoretical computational methods.

Method	Core principle	Advantages	Disadvantages
NIR[4]	Detects overtones and combination vibrations of C–H, O–H, and N–H groups	Identifies functional groups; high penetration	Weak signal/sensitivity; severe spectral overlap
MIR[4]	Excites fundamental vibrations; fingerprint region identifies linkages, anomers, and ring forms	Identifies functional groups; high-resolution linkage and anomer ID	No spatial distribution info.; sensitive to humidity and CO_2 interference
Raman Spectroscopy[4]	Inelastic light scattering via energy exchange/frequency shifts	Identifies functional groups; *in situ* analysis;	Fluorescence/impurity interference; limited spatial resolution
Hyperspectral Imaging[4]	Merged spectroscopy and imaging	Synchronized chemical and spatial mapping	High data redundancy; indirect structural prediction
Methylation Analysis (GC–MS)[1]	Hydroxyl per-methylation → hydrolysis → PMAA identification.	Determines composition, linkage, and branching; widely applicable to plant/fungal glycans	No anomeric (α/β) or sequence info.; tedious and error-prone
Periodate Oxidation[1]	Selective C–C cleavage of vicinal diols	Preliminary linkage/branching assessment	Limited to simple structures; low accuracy for complex glycans
Smith Degradation[1]	Oxidation → rightarrow → reduction → mild hydrolysis	Resolves backbone and repeating units; suitable for complex frameworks	Multi-step and long cycle time; may destroy sensitive motifs
HPLC/UHPLC/HPAEC[2]	Polarity/molecular weight-based separation	High-throughput purity and molecular weight analysis	Lacks fine structural detail
HPSEC-MALLS[2]	SEC separation with laser light scattering	Accurate absolute molecular weight without standards	Only measures molecular weight; no structural unit data
MALDI-TOF-MS[3]	Matrix-assisted laser desorption ionization; TOF analysis of *m/z*.	Intact molecular weight for large glycoconjugates; rapid composition mapping;	Isomer ambiguity; poor ionization for neutral/acidic sugars
LC–MS/MS (ESI)[3]	Electrospray ionization with tandem; MS fragmentation	Resolves sequences, linkages, and modifications; enables quantification	Challenging isomer distinction; complex data interpretation
IM–MS[3]	Separation by ion size/shape	Effectively distinguishes structural isomers	Specialized equipment; limited peak capacity
Nanopore[9]	Decodes structure (sequence, branching) via blockade signatures	Label-free, single-molecule resolution of sequences and modifications in native states	Limited isomer specificity and rapid kinetics; specialized method
X-ray Crystallography[5]	Diffraction-based electron density mapping of single crystals	Atomic resolution; precise conformational (Φ, ψ) mapping	Requires high-quality single crystals; lacks dynamic solution info.
MicroED[5]	Electron diffraction of nanocrystals	Atomic resolution; precise conformational (Φ, ψ) mapping	Restricted to crystalline states; lacks dynamic solution info.
AFM[6]	Tip-sample mechanical interaction	Near-physiological/liquid imaging; visualizes hierarchical assembly	Lower lateral resolution than electron microscopy; no molecular info.
ESI-STM[6]	Soft-landing visualization on surfaces	Single-molecule-resolution	Vacuum/surface environment only; extremely demanding prep.
SEM[7]	Electron beam scanning for surface morphology	Wide field for micro-morphology and sizing	Low resolution; no molecular info.; coating may alter structures
Cryo-EM[7]	Flash-freezing for 3D reconstruction of large complexes	Near-native atomic details for large complexes; no crystallization required	Difficult for flexible/amorphous glycans
MD[8]	Force field-based calculation of conformational space	Constraints for macromolecular polysaccharides conformations; low cost; provides mechanistic insights	Force field dependent; high computational cost for large scales; constraints for macromolecular polysaccharides
NMR[4]	1D/2D/3D pulse sequences (NOE, chemical shifts) for solution-state analysis	Direct solution dynamics and stereochemistry	Signal overlap and low sensitivity for complex glycoproteins; provides ensemble-averaged results

Note: numerical superscripts denote the classification of analytical techniques: 1. Chemical; 2. Chromatography; 3. Mass; 4. Spectrometry, 5. Crystallography; 6. Scanning Probe Analysis; 7. Electron Microscopy; 8. Computational Methods; 9. Electrical and Electrochemical Methods.

The sequence architecture of glycans is highly complex. Monosaccharides can form multiple linkage isomers through glycosidic bonds, while the stereochemistry of the anomeric carbon generates α- and β-anomers. In addition, derivatization further expands structural diversity [1]. Although spectroscopic and chromatographic methods provide information on functional groups and molecular weight, they are insufficient for resolving polysaccharide sequences [1,5]. Therefore, fine-structure elucidation usually relies on the combined use of methylation analysis, selective cleavage methods, NMR, and MS. Methylation analysis and selective cleavage can reveal monosaccharide composition, linkage positions, branching features, and polymerization degree [1]. NMR provides key information on monosaccharide type, substituents, and anomeric configuration, while two-dimensional NMR supports signal assignment, linkage identification, sequence analysis, and isomer differentiation. However, severe signal overlap in highly branched glycans often limits its resolving power [2]. MS platforms, including MALDI–MS and LC–MS/MS, are powerful tools for determining molecular weight, monosaccharide composition, partial sequences, and certain linkage features, yet conventional MS remains limited in distinguishing structural isomers with identical *m/z* values [7,9,10]. Although IM–MS improves isomer separation, its application is constrained by finite peak capacity and limited accessibility [9]. Nanopore sensing has emerged

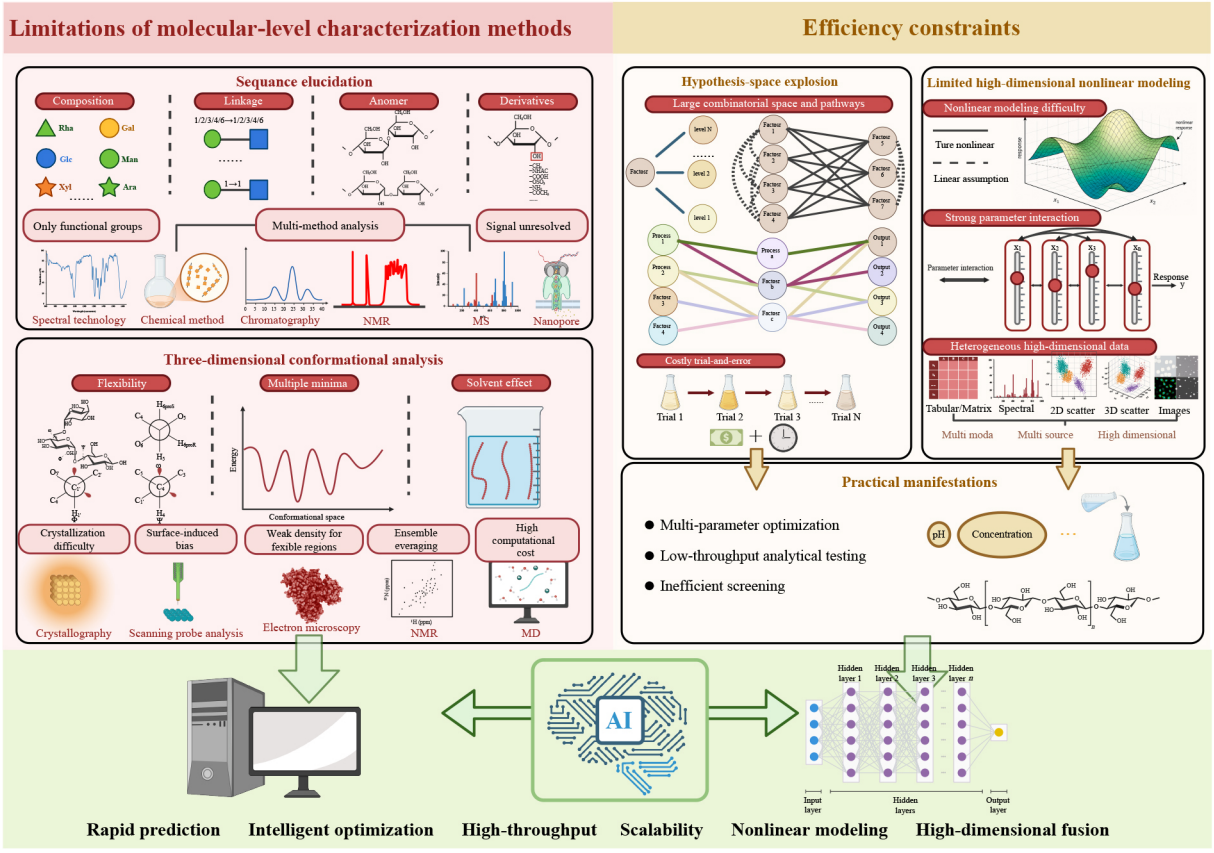


Fig. 2 Schematic diagram of existing problems in the field of polysaccharides.

as a promising strategy for single-molecule glycan sequencing, but its practical use is still restricted by low sensitivity, insufficient isomeric resolution, and challenges related to translocation kinetics and sample purity [11].

Glycans also exhibit highly flexible 3D structures, with low energy barriers between conformations and rapid interconversion among conformers. AFM and SEM provide morphological information at the nano- and microscale, but higher-resolution techniques are required for atomic-level insight [12]. Current approaches for glycan 3D structural analysis include crystallography, NMR, cryo-electron microscopy (cryo-EM), electrospray ionization scanning tunneling microscopy (ESI-STM), and molecular dynamics (MD) simulations. Crystallography has resolved lectin-glycan complexes, glycopeptides, and polysaccharide crystals, but it often fails to preserve solution-state features and is limited by crystal heterogeneity and resolution [13]. NMR remains a major tool for glycan conformational analysis, although signal overlap is still problematic for oligosaccharides and polysaccharides with repeating units [14]. ESI-STM enables single-molecule visualization without ensemble averaging [12], but its dependence on surface states restricts the detection of low-abundance solution conformations. Cryo-EM has revolutionized structural biology [15], yet glycan analysis remains challenging because rapid vitrification captures multiple conformations of flexible sugar chains and associated protein regions, complicating image alignment and limiting glycan resolution [16,17]. Computational chemistry, including quantum chemical calculations and molecular mechanics methods such as GLYCAM06, complements experimental techniques [18]. For example, conformational libraries generated by MD simulations can be integrated with cryo-EM data to accelerate glycoprotein structure modeling [16]. However, the computational cost rises sharply with atom number and polymerization degree, limiting the large-scale application of these methods in glycoscience.

In summary, the substantial diversity of polysaccharide sequences and three-dimensional conformations imposes inherent limitations on current methods for sequence determination and structural elucidation. Even when multiple complementary analytical techniques are combined, complete and unambiguous characterization of polysaccharide structures remains difficult. Moreover, experimental errors and differences in sample preparation, instrumental conditions, data-processing procedures, and analytical standards reduce the comparability and reproducibility of results across studies, thereby limiting data integration and the establishment of reliable

molecular-level structural models. Inspired by the transformative success of AlphaFold in protein structure prediction [8], ML has shown strong potential in the study of bioactive macromolecules. Therefore, introducing ML-based strategies into glycan structural elucidation is urgently needed to overcome the limitations of conventional methods and promote advances in polysaccharide science and its applications.

2.2 Efficiency constraints

Due to the lack of molecular-level theoretical support, traditional research methods face massive combinatorial search spaces generated during hypothesis formation and validation. This not only dramatically increases time and economic costs but also falls far short in modeling capabilities for high-dimensional non-linear systems, rendering the entire process beyond the effective scope of conventional analytical methods. This dilemma is particularly prominent in polysaccharide research (Fig. 2). On one hand, polysaccharide extraction processes involve core parameters such as solid–liquid ratio, temperature, pH, and extraction time. The introduction of various auxiliary technologies—such as enzymatic digestion, acid-base treatments, ultrasound, supercritical CO_2 extraction, deep eutectic solvents, ionic liquids, and three-phase solvent systems [19–26]—further exponentially increases the dimensionality and complexity of the parameters affecting extraction efficiency. The preparation of glycan-based materials requires balancing multiple factors, including principle selection, component ratios, and processing conditions, while simultaneously meeting diverse optimization objectives such as mechanical properties, functional characteristics, and biological activities [3,27–29]. Traditional response surface methodology generally identifies local optima within a restricted experimental region rather than guaranteeing global optimality. Moreover, as the number of experimental factors increases, the required number of design points grows substantially, making global optimization increasingly challenging [26]. On the other hand, traditional quality testing of food and agricultural products relies heavily on chemical analysis. These methods universally suffer from tedious sample preparation, destructive testing processes, high reliance on professional operators, expensive instruments and consumables, and low testing efficiency. They are also susceptible to environmental interference and sample contamination, leading to poor result stability. Furthermore, they can only acquire single-point holistic data for a sample, failing to visually characterize physical features, spatial distribution, and local defects, making them unsuitable for the high-throughput, rapid quality testing demands of the food industry [30]. Furthermore, the evaluation of glycan bioactivity still depends largely on cell- and animal-based experiments. In the absence of critical molecular-level information, precise activity characterization and mechanistic interpretation remain difficult to achieve using conventional approaches. In addition, glycan biomarker discovery is often challenged by the massive volume, high dimensionality, and structural heterogeneity of glycomics data, which are difficult to analyze efficiently using traditional methods. Given these limitations, ML offers a powerful alternative because of its strong capability in high-dimensional data analysis and nonlinear relationship modeling.

3 Model Training

3.1 Data preparation

The foundational step in constructing ML models is the assembly and curation of datasets for model training. The data types covered in this review encompass component ratios, process parameters, spectral data, imaging data, glycan structural data, macroscopic physicochemical properties, and biological activity data. The second step in model building is typically data preprocessing, which is crucial for enhancing model quality and reliability; this primarily includes data cleaning, feature selection, feature scaling, and dataset partitioning. Currently, datasets used for model training are mostly derived from experimental measurements. However, due to inconsistencies in experimental methods, testing platforms, and data standards, datasets broadly suffer from significant heterogeneity, chaotic formatting, small scale, and inconsistent annotations. Consequently, this paper only summarizes and compiles large-scale datasets that have undergone standardized curation (Table 2).

3.2 ML algorithms

ML can generally be categorized into supervised, unsupervised, and semi-supervised or reinforcement learning. Supervised learning uses labeled data to build predictive models, whereas unsupervised learning aims to uncover latent patterns from unlabeled data. As shown in Table 3, the algorithms discussed in this review are broadly classified into traditional ML algorithms, DL algorithms, and optimization algorithms. Traditional ML includes Random Forest (RF), Support Vector Regression (SVR), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Extreme Gradient Boosting (XGBoost), Partial Least Squares Regression (PLSR), and k-Nearest Neighbors (KNN). DL includes Artificial Neural Networks (ANN), Probabilistic Neural Networks

Table 2 Glycan-related databases.

Name	Data type description	Data volume
GlyTouCan	A registry covering fundamental monosaccharide composition to high-resolution, fully defined complex glycan topologies and linear sequence data	Over 240,000 registered structures
CSDB	3D/topological structures, NMR spectroscopy data, and phylogenetic/literature metadata (focusing on prokaryotes, archaea, fungi, and plants)	~34,000 natural carbohydrates and glycoconjugates
KEGG GLYCAN	Experimentally determined structures and glycan network data deeply mapped into biochemical metabolic networks and biosynthesis pathways	11,251 items
UniCarb-DB/UniCarbKB	Focuses on experimental LC–MS/MS data and annotations for N- and O-linked glycans released from glycoproteins	> 1588 spectra and 1118 defined structures
GlycoPOST	Raw MS data repository for glycomics and glycoproteomics	563 registered projects, ~6.6 TB data
GlycoNMR	Precise carbohydrate 3D structures and exhaustively annotated NMR atomic-level chemical shift signature data	2609 3D structures, 211,543 annotated chemical shifts
DIONYSUS	Carbohydrate molecular fragments in protein-interacting states and their high-resolution atomic-level 3D interface spatial feature data	> 330,000 molecular fragments, ~50,000 experimental structures
CAZy	Biological features and structural data of enzyme families responsible for synthesizing, degrading, or modifying complex glycosidic bonds	Tracks and includes > 36,000 cross-species enzyme distributions
SugarBindDB	Specific binding pair events and ligand information between pathogenic microorganism lectins and human host-targeted glycans	1266 validated specific binding events
Glyco3D	3D spatial conformations and NMR spectral data of mono/di/polysaccharides, glycosyltransferases, antibodies, etc.	Covers various 3D conformational records (e.g., ~150 entries for monosaccharides)
glycowork	Glycan sequences (with species/tissue/disease labels) and out-of-the-box pre-trained records for carbohydrate-protein binding interactions	~50,500 independent glycan sequences, >790,000 binding records
GlyGen	Federated multi-omics data integrating and standardizing various glycans, glycoconjugates, proteomes, and genomic mutation association data	Large-scale interdisciplinary federated data collection
GlyCosmos	Interdisciplinary multi-omics portal deeply mapping glycan structures to upstream genomic/proteomic and downstream lipidomic pathways	Includes 5100 lectin entries and 6046 glycolipid entries, etc.
CAS	Registered records of precise chemical structures and physical properties for glycan monomers and complex carbohydrate-containing molecules	> 1.16 million structure records containing glycan monomers
GlycoShape	3D structures of human and other biological glycans derived from cumulative MD simulation sampling	> 500 unique glycan 3D structures
Consortium for Functional Glycomics	Glycan microarray experimental binding affinity data for mammalian/microbial glycans interacting with various glycan-binding proteins	5203 glycan microarray sample data points
Protein Data Bank	Contains independent oligosaccharide structures using branched entity representation and 3D crystal structures of protein-carbohydrate complexes	> 8000 oligosaccharide structures and thousands of complex structures
MassIVE	Raw experimental MS and quantitative omics data for glycomics and glycopeptides from human and model organisms	> 50,000 MS datasets containing massive glycomics-specific datasets
NIST (NIST Glycan Mass Spectral Library)	Gold-standard reference MS/MS spectra for core glycans, human milk oligosaccharides (HMOs), and monoclonal antibody glycopeptides	16,421 mass spectra covering 209 core glycans

(PNN), Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), Graph Convolutional Networks (GCN), Transformers, Dilated Residual Networks (DRN), and Convolutional Neural Networks (CNN). Optimization algorithms include Genetic Algorithm (GA), Non-dominated Sorting Genetic Algorithm II (NSGA-II), Particle Swarm Optimization (PSO), Differential Evolution (DE), Starfish Optimization Algorithm (SFOA), Whale Optimization Algorithm (WOA), and Sparrow Search Algorithm (SSA).

3.3 Evaluation

Before applying models to predictive tasks, standard evaluation protocols must guide model optimization. Hold-out validation, cross-validation, and bootstrapping are mainstream evaluation schemes. For classification models, multidimensional metrics are typically used to assess performance comprehensively, including accuracy, recall, specificity, precision, F1-score, receiver operating characteristic (ROC) curve, and area under the curve (AUC). The ROC curve and AUC primarily measure the overall discriminative ability of the model; an ROC curve closer to the top-left corner and an AUC value closer to 1 indicate superior overall performance. For regression models, common evaluation metrics include mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), ratio of performance to deviation (RPD) and the coefficient of determination (R^2).

4 From Preparation and Structural Elucidation to Rational Reverse Design

This section summarizes the current research status of ML in the investigation of polysaccharide molecular structures and highlights the novel paradigm enabled by ML for this research field. As shown in Fig. 3, the review covers polysaccharide extraction and synthesis, sequence analysis, 3D structures prediction, interaction prediction and *de novo* design.

4.1 ML-optimized glycan extraction and synthesis

4.1.1 Extraction

Polysaccharides are generally classified into starch and non-starch polysaccharides, the latter of which can be

Table 3 Principles, advantages, and disadvantages of ML algorithms.

Algorithm	Principle	Advantages	Disadvantages	Reference
RF	Bagged decision-tree ensemble with random feature selection	Robust against overfitting; handles non-linear relationships and high-dimensional data	Limited interpretability; biased feature importance; weak extrapolation	[31,32]
SVR	Regularized regression using an ε-insensitive loss and kernels	Effective for nonlinear, high-dimensional, small-to-medium datasets	Sensitive to scaling, kernel choice, and hyperparameters; costly for large datasets	[33]
SVM	Maximum-margin classification with optional kernel mapping	Effective in high-dimensional and small-to-medium datasets	Sensitive to kernel and regularization parameters; limited scalability	[34]
LDA	Linear projection maximizing between-class relative to within-class separation	Fast, interpretable, and useful for dimensionality reduction	Assumes linear separation and similar covariance; unstable in high-dimensional small samples	[35]
XGBoost	Regularized gradient boosting of decision trees	Strong nonlinear modeling; handles interactions, missing values, and sparse data	Requires careful tuning; may overfit; limited interpretability	[36,37]
PLSR	Latent-variable regression maximizing predictor-response covariance	Suitable for collinear and high-dimensional spectral data	Mainly linear; sensitive to outliers and latent-variable selection	[36,38]
KNN	Prediction based on the *k* nearest training samples.	Simple, nonparametric, and able to capture local patterns	Sensitive to scaling, distance metric, and dimensionality; slow prediction on large datasets	[39]
ANN	Layered nonlinear mapping trained by backpropagation	Flexible for complex classification and regression tasks	Prone to overfitting in small datasets; tuning and interpretation are difficult	[40,41]
PNN	Kernel density estimation combined with Bayesian classification	Rapid training and nonlinear probabilistic classification	Sensitive to smoothing parameters; memory and prediction costs increase with sample size	[42]
LSTM	Gated memory cells retain and update sequential information	Captures long-term temporal dependencies better than conventional RNNs	Computationally intensive; requires meaningful sequential data and careful tuning	[43,44]
RNN	Recurrent hidden states transmit previous sequential information	Suitable for temporal and variable-length data	Vanishing or exploding gradients hinder long-term learning; limited parallelism	[45,46]
GCN	Aggregates information from neighboring nodes in a graph	Suitable for molecular graphs and relational data	Depends on graph quality; deep networks may suffer from over-smoothing	[47]
Transformer	Multi-head self-attention models global sequence relationships	Captures long-range dependencies and enables parallel training	Quadratic attention cost; high data and computational requirements	[48,49]
DRN	Residual networks with dilated convolutions enlarge the receptive field	Preserves resolution while capturing multiscale features	May produce gridding artifacts; architecture requires careful design	[50]
CNN	Shared convolutional filters extract hierarchical local features	Efficient for image, spectral, and spatial-pattern recognition	Requires adequate labeled data; pooling may lose details; limited global context	[51,52]
GA	Population evolution through selection, crossover, and mutation	Gradient-free and applicable to complex search spaces	Sensitive to operators and population settings; may converge prematurely; no global-optimum guarantee	[53,54]
NSGA-II	Elitist non-dominated sorting with crowding-distance preservation	Produces diverse trade-off solutions for multi-objective problems.	Only approximates the Pareto front; depends on population and operator settings	[55]
PSO	Particles move according to personal and swarm best positions	Simple, gradient-free, and requires few main parameters	May stagnate or converge prematurely; sensitive to update coefficients	[56,57]
DE	Population search using differential mutation, crossover, and selection	Simple and effective for continuous nonlinear optimization	Sensitive to mutation, crossover, and population settings; no global-optimum guarantee	[58]
SFOA	Starfish-inspired exploration, predation, and regeneration operators	Balances global exploration and local exploitation	Recent method with limited independent and food-related validation	[59]
WOA	Simulates prey encircling, spiral feeding, and random search	Simple structure combining exploration and exploitation	May lose diversity or become trapped in local optima	[60,61]
SSA	Producer-scrounger and warning mechanisms simulate sparrow foraging	Simple structure and rapid convergence in some problems	May show insufficient local search and premature convergence	[61,62]

further divided into microbial and non-microbial types. ML algorithms can optimize the extraction processes of all these polysaccharides and enable single- or multi-objective process control guided by target requirements such as glycan structure and biological activity.

In process optimization, extraction yield is usually the primary goal. Researchers typically begin by identifying key variables affecting extraction efficiency using methods such as Plackett–Burman (PB) design, gray relational analysis (GRA), or Least Absolute Shrinkage and Selection Operator (LASSO) regression [22,63,64]. Based on response surface designs, experimental data are then used to establish predictive models linking process parameters to extraction performance, with algorithms such as RF, XGBoost, ANN, and SVR widely applied [22,64,65]. Global optimization methods, particularly genetic algorithms (GA), are subsequently employed to determine the optimal conditions for maximum yield [26]. Through this workflow, the extraction yields of both starch and non-starch polysaccharides can be significantly improved [22,66–68]. ML has also been applied to optimize green extraction technologies, including enzyme-assisted extraction [22], ultrasound-assisted extraction [69], supercritical CO_2 extraction [19], deep eutectic solvent extraction [70], ionic liquid extraction [24], and three-phase partitioning extraction [26].

However, the optimal conditions identified by these models are often specific to the raw material and experimental batch used for model development. Natural feedstocks can vary substantially with geographical origin, harvest period, cultivar, storage history, and batch, which may change their composition and extraction

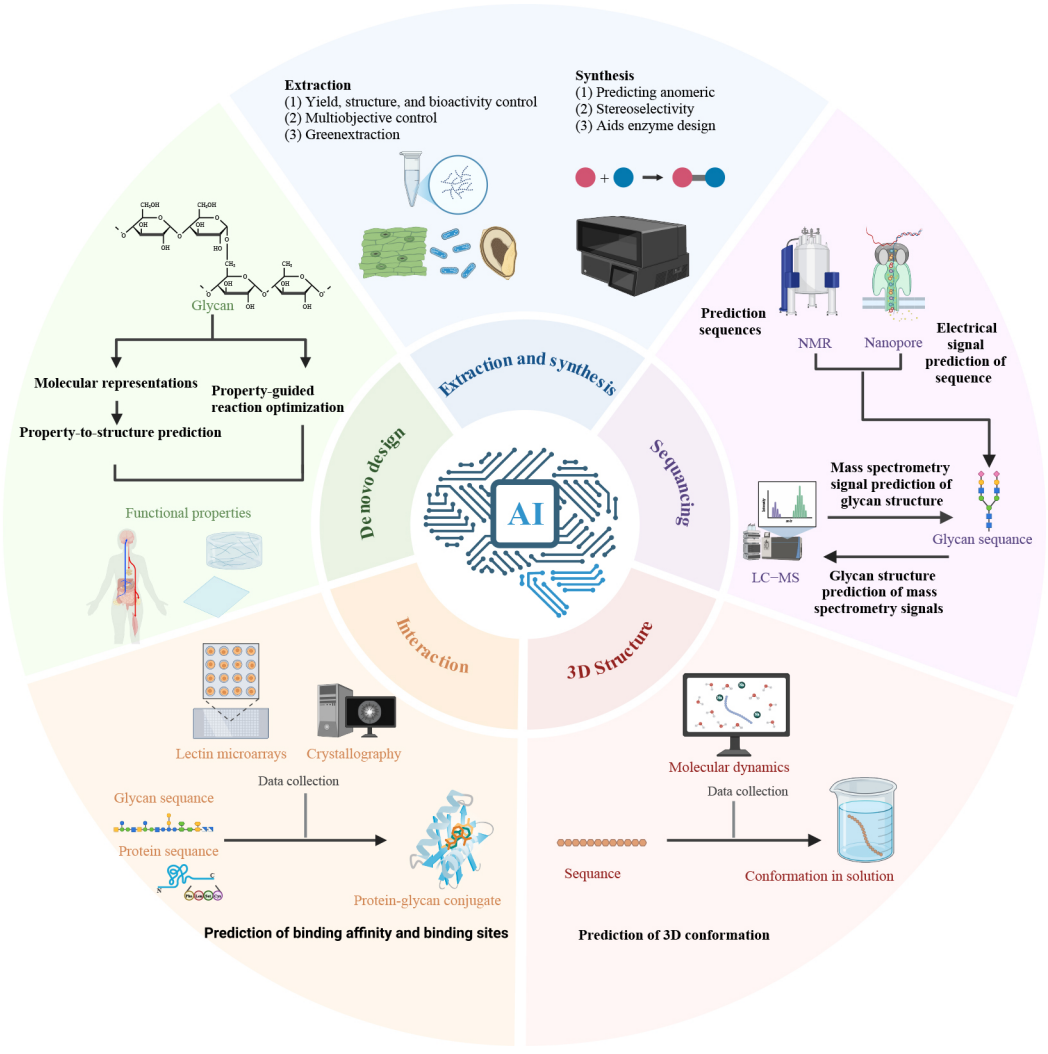


Fig. 3 Machine learning-assisted polysaccharide research workflow from extraction to *de novo* molecular design.

response. Therefore, conditions optimized from a single batch should not be assumed to transfer directly to practical applications; external validation across independent sources, harvest periods, and batches is needed to assess robustness and generalizability.

Given that extraction methods impact polysaccharide structure, and that structural integrity and biological activity are widespread concerns, ML can also optimize processes and screen extraction methods based on structural requirements and activity orientation. For instance, using parameters like the polygalacturonase/substrate ratio, cellulase/polygalacturonase activity ratio, and reaction time as inputs, single-output prediction models have been constructed for 10 indicators (e.g., solid yield of sugar beet pulp enzymatic hydrolysates, polysaccharide-to-mono/oligosaccharide conversion efficiency), enabling structure-oriented precise process optimization [71]. Other studies have used structural descriptors, such as molecular weight, monosaccharide composition, and branching degree, to distinguish different extraction methods and support intelligent method selection [72]. In addition, neural network models linking process parameters with antioxidant activity and extraction yield have enabled activity-guided targeted optimization [20].

Notably, polysaccharide extraction is inherently a multi-objective optimization problem requiring the simultaneous consideration of purity, structure, function, bioactivity, production cost, and extraction efficiency. Unlike single-objective optimization, such tasks are typically addressed through Pareto-optimal solution sets, which can be effectively optimized using algorithms such as NSGA-II. For instance, multi-output models combined with NSGA-II have been used to simultaneously improve total sugar content, dry extract yield, and the yields of specific active compounds [63]. Overall, ML enables

requirement-driven process control, intelligent extraction method selection, and efficient multi-objective optimization in polysaccharide extraction. With continued advances in glycomics, future studies may achieve more precise optimization targeting glycan sequence features and ultimately establish more generalizable models for diverse polysaccharide sources.

4.1.2 Synthesis

The chemical synthesis of polysaccharides is a critical foundation for conformational analysis and the development of carbohydrate-based functional materials, serving as a key step in programmable glycan assembly. Although automated glycan assembly technology has successfully synthesized a 1080-mer glycan (the longest homogeneous polysaccharide reported to date), precisely controlling the anomeric carbon stereochemistry during glycosylation remains a core challenge hindering efficient synthesis [12].

Recently, the application of ML in reaction prediction and enzyme design has offered novel strategies to overcome the anomeric stereoselectivity challenge. Researchers screened and integrated 10,434 single-step glycosylation reactions from the CAS database (2010–2015) to build GlycoPredictor the largest, top-performing, and only publicly accessible ML prediction model for glycosylation reactions globally. This model accurately predicts the dominant anomeric product and its ratio based on key parameters like glycosyl donor, acceptor, solvent, activator, and reaction temperature [73]. Concurrently, the rapid advancement of ML in enzyme design [74] opens new pathways for the efficient and precise regulation of anomeric selectivity.

4.2 ML-driven glycan sequence identification

As previously noted, glycan sequence structures are highly complex, encompassing monosaccharide composition, anomeric configurations, branching sites, and chain sequences, usually requiring a combination of MS, NMR, and methylation analysis for complete sequencing. To achieve efficient and accurate polysaccharide sequencing, researchers have adopted end-to-end strategies, utilizing ML algorithms to intelligently identify and elucidate glycan sequences based on NMR data, nanopore current signals, and MS data.

4.2.1 Based on NMR and Nanopores

ML-assisted NMR data analysis has realized the sequence identification of di- and tri-saccharides [75]. However, NMR-based structural elucidation of polysaccharides continues to face challenges: severe signal overlap (especially in high-field regions) and spectral complexity caused by conformational flexibility. The analysis of larger glycan structures remains limited; nevertheless, considering the learning capabilities of LSTM and RNN models on multi-dimensional NMR spectra, breakthroughs are anticipated [4].

Nanopores can reversibly bind with glycan molecules, generating specific, stable, and distinguishable characteristic ionic current signals. However, the current signals produced as glycans traverse nanopores are typically weak, complex, and highly overlapping, making manual discrimination exceptionally difficult. ML technology can automatically extract key features such as current blockade amplitude, dwell time, and standard deviation, enabling high-resolution identification at the monosaccharide level [11]. For example, ML can effectively distinguish maltose oligosaccharides with degrees of polymerization ranging from 1 to 8 [76], as well as glycosaminoglycan disaccharides with varying monosaccharide compositions [77]. Based on the three major nanopore glycan sequencing strategies currently proposed (strand sequencing, sequential hydrolysis sequencing, and assembly sequencing) [11], nanopore technology is poised to achieve high-precision structural elucidation and sequencing of large, complex glycans.

4.2.2 Based on MS

Early glycan sequencing methods relied on searching and comparing MS data against databases, which was restricted by missing structural signature ions and the inability to handle structural heterogeneity [78]. Recently, MS data-driven ML prediction algorithms aim to correct high-throughput glycan sequencing. ML methods have massive data requirements; inevitably, related reports are skewed toward the most abundantly generated glycan data, which are often N-/O-linked glycans and those associated with glycan-binding proteins.

Current ML-based glycan sequencing research mainly follows two directions. The first predicts MS data from glycan structures. DeepGlyco uses glycan structure encoding, tree-LSTM feature extraction, fragmentation-site attention, and ion intensity prediction to achieve highly accurate MS peak prediction, with spectral angle loss below 0.11 and dot product above 0.98, while also distinguishing structural isomers [79]. DeepGP, a hybrid framework combining Transformers and GCNs, predicts intact N-glycopeptide MS/MS spectra and retention times, improves glycopeptide identification sensitivity, and enables structure identification even without diagnostic ions, with performance exceeding 90% across multiple datasets [78]. Nevertheless, both methods still have limited applicability, with DeepGlyco restricted to

glycans composed of five specific monosaccharides [79], while DeepGP is designed only for N-glycopeptides and does not account for the influence of anomeric isomers on MS behavior [78].

The second direction infers glycan structures directly from MS data. GlycanFinder performs *de novo* sequencing of N-linked glycans by combining dynamic programming, Graphormer-based GNNs, and ion-matching neural networks, but its reported accuracy remains relatively low at 32% [80]. In contrast, CandyCrunch represents a major advance. Built on the largest LC–MS/MS dataset for eukaryotic glycans reported to date, containing 489,103 validated spectra, CandyCrunch integrates a dilated residual neural network with biosynthetic priors and GNN-based topology modeling to enable fragment-ion assignment, sequence inference, and biosynthetic pathway deduction in a unified framework. It can rapidly predict and classify 3391 known glycans from raw LC–MS/MS data, achieving an overall structural recognition accuracy of 90.3% (F1-score = 0.897) [10].

Earlier ML-assisted glycan sequencing studies were generally unable to resolve anomeric configuration, which limited glycan classification, monosaccharide discrimination, and overall sequencing accuracy [78,80]. CandyCrunch addressed several of these limitations by achieving anomeric configuration prediction, expanding the scale of training data, and introducing biosynthetic network constraints to reduce biologically implausible false positives [10]. Despite these advances, high-throughput and high-precision prediction of glycans from a broader range of biological sources still requires further investigation.

4.3 Three-dimensional structure

The 3D structure of biomolecules fundamentally determines their physicochemical properties and biological functions, which ultimately arise from specific intermolecular interactions. Owing to the limitations of traditional analytical techniques, progress in glycan 3D structural elucidation and molecular interaction studies has lagged far behind that of proteins and nucleic acids, despite glycans being among the most widely distributed biological macromolecules in nature [12]. ML-based structural prediction methods, exemplified by AlphaFold3, have greatly advanced protein structure determination, functional annotation, and interaction analysis [8]. With the accumulation of crystallographic data and the continued development of MD simulations, glycan research is now entering a new stage in which ML offers promising opportunities for 3D conformation prediction, molecular recognition analysis, and structure-function investigation.

Recent ML models for glycan 3D structure prediction are largely built on MD-derived conformational databases such as GlycoShape [81]. For glycans with low degrees of polymerization, all-atom MD can effectively characterize structural and dynamic features because their glycosidic bonds and conformational degrees of freedom are relatively manageable [82,83]. GlycoShape has established a database of 534 glycans based on a cumulative 1 ms of MD simulations, with representative conformations identified by clustering analysis [84]. Based on GlycoShape, Luc Thomès et al. [85] developed a deep learning framework that takes glycan sequences as input and combines Graph Isomorphism Networks with von Mises mixture distributions to model the multimodal distributions of glycosidic torsion angles. The model predicts torsion-angle distributions, monosaccharide-level solvent-accessible surface area, and flexibility values directly from sequence. For previously unseen glycans, it achieved prediction errors of only $\varphi \pm 7°$, $\psi \pm 9°$, and $\omega \pm 10°$, with a Pearson correlation coefficient of 0.94 for solvent-accessible surface area. This work represents the first deep learning approach capable of predicting glycan 3D structural properties directly from sequence alone. Overall, ML-based glycan 3D structure prediction has progressed rapidly in recent years, but accurate prediction for large glycans and polysaccharides remains challenging. Given the high conformational similarity among repeating units in polymeric glycans [14], inferring global conformations from representative repeating-unit structures may provide a promising route to overcome this limitation.

4.4 Interactions

Current studies of glycan interactions with other substances are focused mainly on glycan-protein interactions, with the major tasks being binding affinity prediction and binding site prediction. For binding affinity prediction, most models use glycan sequence information as the core input to quantitatively estimate glycan-protein binding strength. A representative example is a dual-encoder framework that combines LSTM-based viral protein sequence features with SweetNet-derived glycan graph features to predict binding affinity *Z*-scores, achieving an optimal MSE of 0.735 on 120,000 interaction records and successfully identifying glycan receptors for influenza viruses, coronaviruses, and rotaviruses [86]. GlyNet encodes glycan sequences into 272-dimensional q-gram features to predict continuous RFU values for 1257 protein samples, showing strong performance (MSE = 0.12, R^2 = 0.78, AUC = 0.912) and enabling extrapolation to 4160 novel glycans [87]. However, its predictive scope is restricted to a fixed set

of proteins. More recent studies have shown that model performance depends strongly on glycan training set coverage, molecular representation, and structural feature incorporation. MCNet, for example, uses chirality-aware Morgan fingerprints and atomic q-grams, and is trained on both Consortium for Functional Glycomics glycan array data and glycomimetic datasets [88]. It achieved the first successful modeling of cross-chiral glycan recognition and experimentally confirmed the binding of L-glucose to fucose-binding lectins. In addition, incorporating 3D descriptors, including solvent-accessible surface area, flexibility, and torsion angles, into SweetNet improved affinity prediction accuracy by 7.4% [85]. These advances indicate that glycan-protein affinity prediction has evolved from sequence-based modeling toward integrated frameworks that combine sequence, 3D structure, and chirality-related information.

Binding site prediction, in contrast, relies largely on crystallographic data. With the expansion of carbohydrate-protein interaction datasets, this field has progressed from traditional machine learning methods trained on relatively small datasets [89,90] to high-precision deep learning models supported by large-scale structural data. DeepGlycanSite is a representative example, integrating deep equivariant graph neural networks with Transformers to predict glycan-binding pockets and specific glycan-protein binding sites. In this framework, both proteins and glycans are encoded as atomic-level molecular graphs, enabling structural feature extraction and binding site prediction in a unified architecture [91]. Its performance surpasses that of existing tools such as GlycoTorch Vina, AutoDock Vina, DiffDock, and EquiBind. GlycanInsight further extends this framework into a web platform that supports both experimental and AlphaFold2-predicted protein structures, as well as multiple glycan ligand formats [92].

4.5 *De novo* design guided by functional properties

Polysaccharides are widely used as functional matrices in food, biomedicine, and biomaterials, and their macroscopic physicochemical properties, including rheological behavior, gel strength, solubility, and texture, together with their biological activities, directly determine the performance, application value, and market acceptance of final products. Accordingly, the design of polysaccharides and polysaccharide-based composite systems guided by target properties and activities has become an important strategy for precise and functional development. In this review, AI-assisted *de novo* design encompasses two complementary directions: property-to-structure design and reaction-condition-to-property optimization. The latter is included because predictive models can be coupled with optimization algorithms, such as GA, to identify conditions that achieve desired properties. It should also be noted that, because glycan-specific studies remain limited, selected studies from other fields are included as methodological references for future glycan and polysaccharide design.

4.5.1 Property-to-structure design

For property-to-structure prediction, the structure of the target material must be represented as the model output. Polysaccharide structures can be encoded using molecular fingerprints, topological indices, and molecular graph- or string-based representations [84]. For instance, carbohydrate structures encoded as strings have been used for immunogenicity prediction with accuracies above 92% [93]. Glycan molecular graphs have also been employed to predict binding strength between glycans and viral proteins, demonstrating potential for glycan-based antiviral design [86]. In parallel, molecular fingerprint-based approaches have been shown to effectively predict the binding strength between L-conFig.d carbohydrates and proteins [88]. As previously demonstrated, researchers have shown that incorporating stereochemical and spatial structural information can significantly improve predictive accuracy. In addition, topological descriptors derived from molecular graphs can capture spatial structural information of molecules and have been shown to achieve strong predictive performance [94]. Notably, the proposed topological descriptor framework enables the use of polymerization degree as an input feature to generate corresponding quantitative outputs for property prediction. Although amylose is composed of relatively simple linear glucose units linked via α-1,4-glycosidic bonds, this study provides an important methodological insight: partial structural descriptors can be effectively mapped to macroscopic physicochemical properties, offering a generalizable strategy for predicting glycan-related functions from incomplete or abstracted molecular structural information.

These structural representations provide a foundation for property-to-structure design by enabling candidate structures to be represented, generated, and evaluated in a machine-readable space. Building on these representations, the next step is to establish a conceptual framework for AI-assisted property-to-structure design. From a statistical perspective, this process can be theoretically formulated as sampling structures or sequences that satisfy predefined target properties from a conditional distribution, $x \sim p(x \mid y \in Y)$. In this framework, a general generative model, $p(x)$, learns the prior distribution of plausible structures, sequences, or atomic configurations, whereas a property predictor, $p(y \mid x)$, estimates the like-

lihood that a candidate exhibits desired properties, such as stability, activity, or binding affinity. These two components can be combined through Bayes' rule, $p(x|y) \propto p(y|x)p(x)$, thereby enabling inverse design from desired properties to candidate structures or sequences. Depending on how property information is incorporated, AI-assisted property-to-structure design can generally be implemented through three strategies: baked-in conditioning, in which target properties are explicitly incorporated during model training; adaptive conditioning, which combines a general generative model with an independent property predictor; and on-the-fly guidance, in which property information dynamically guides the sampling process during generation [22]. Thus, rather than simply identifying a single candidate with the highest predicted score, property-to-structure design aims to learn a feasible design space and guide the generative distribution toward regions associated with the desired properties, thereby producing multiple candidate structures or sequences that satisfy the predefined design criteria.

4.5.2 Reaction-condition-to-property optimization

For reaction- or formulation-condition-to-property optimization, a study established an iterative ML-assisted framework for the inverse design of underwater adhesive hydrogels. Based on an initial dataset of 180 bioinspired hydrogels, the relative proportions of six functional monomers were used as input variables and underwater adhesive strength as the target property. Gaussian process and random forest regression models were first trained to predict adhesive strength, and were subsequently coupled with batched sequential model-based optimization to propose new monomer compositions with improved performance. Through three iterative rounds of ML prediction, experimental synthesis, property measurement, and model updating, the dataset was expanded to 341 formulations and the optimal composition was progressively identified, ultimately yielding a hydrogel with an underwater adhesive strength exceeding 1 MPa [95]. With advancements in glycan chemistry, the precise design of glycan molecules has gradually become feasible [12]. Therefore, integrating glycan chemical synthesis with ML algorithms to design product macroscopic properties and biological activities is emerging as a core paradigm for realizing the *de novo* design of glycans and glycan-based systems.

5 Applications of ML-Driven Polysaccharide Research

Polysaccharides serve as one of the essential material carriers of biological information and have been widely applied in the food industry. This section reviews the applications of ML in biomarker and bioactivity screening, quantitative identification and traceability of polysaccharides in agricultural products and processed foods, as well as material fabrication as illustrated in Fig. 4.

5.1 Biomarker discovery and biological activity screening

5.1.1 Biomarker discovery

The human glycome encompasses N-, O-, and C-linked glycans, glycosphingolipids, proteoglycans, glycosaminoglycans, and exogenous glycans. Under pathological conditions, aberrant glycan biosynthesis and glycosylation pathways serve as direct indicators of disease states [96]. Recently, the synergy between ML and glycomics has facilitated the identification of multivariate glycan biomarkers, overcoming the limitations of traditional single-molecule markers. Adhering to the biomarkers, endpoint, and other tools (BEST) criteria framework, ML-driven approaches have identified diagnostic, predictive, prognostic, and susceptibility glycan biomarkers across diverse pathologies, including malignancies (digestive, respiratory, and urogenital systems), cardiovascular diseases, autoimmune disorders, and chronic metabolic conditions such as chronic kidney disease and fatty liver [96–103].

In clinical diagnostics and therapeutic response prediction, these novel biomarkers demonstrate superior performance. Diagnostic ML models enable early-stage screening (e.g., detection of Stage I-II colorectal cancer [98] and Stage 0-II non-small-cell lung cancer [97], facilitate differential diagnosis in complex pathological contexts (e.g., distinguishing hepatocellular carcinoma from cirrhosis or chronic hepatitis B [100]), and achieve precise subtyping (e.g., lung adenocarcinoma vs. squamous cell carcinoma [97]; infectious vs. non-infectious febrile states [104]). Furthermore, predictive models facilitate the early identification of clinical responses to specific interventions, such as chemotherapy [101] or surgery [105], providing a critical foundation for individualized targeted therapy. Additionally, ML is increasingly applied to prognostic quantification and latent risk assessment. Prognostic models categorize disease trajectories, such as overall survival and recurrence-free survival in oncological patients [102]. Regarding risk assessment, glycan biomarkers are utilized to estimate "biological age". By quantifying organ-specific glycosylation age (e.g., renal GlyKage) and its acceleration (GlyKageDiff) [103] researchers can digitally assess latent cardiovascular risk and organ senescence.

ML-based multimodal biomarker discovery effectively integrates multidimensional biological data to charac-

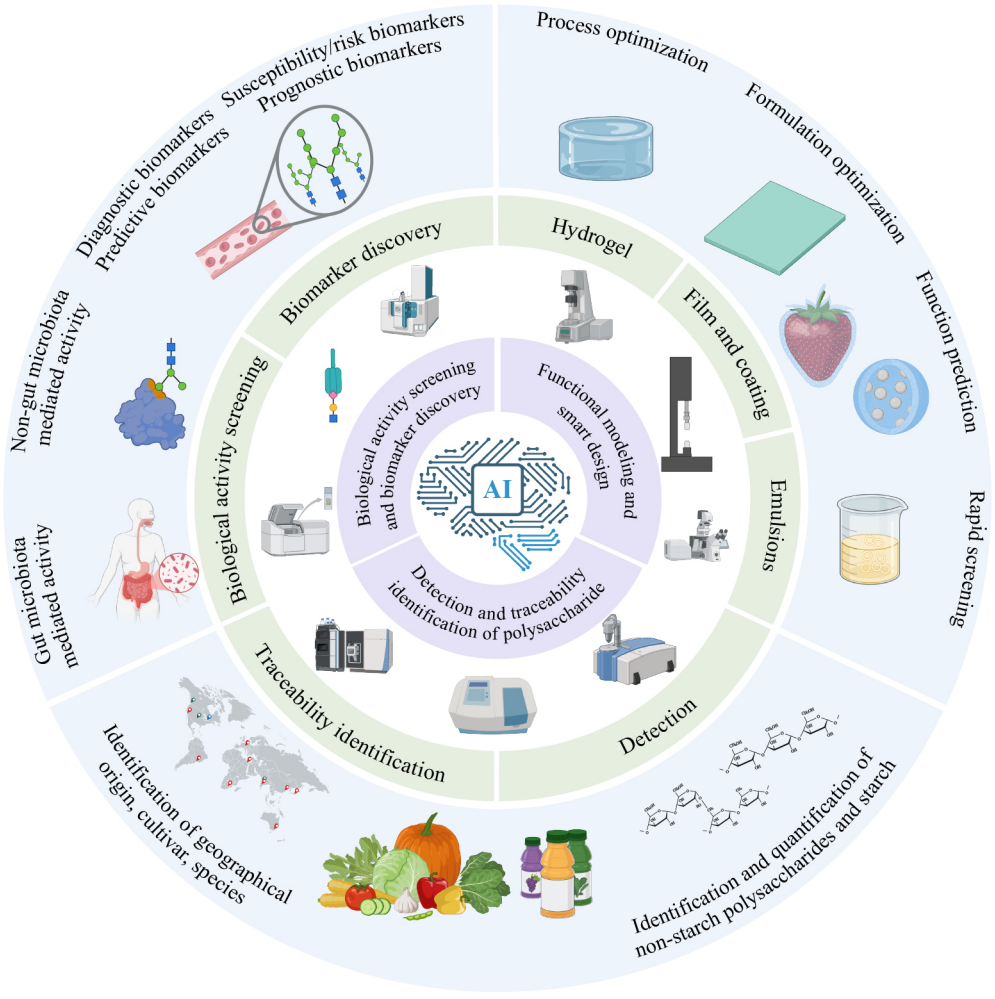


Fig. 4 Applications of machine learning in polysaccharide characterization, prediction, and intelligent material design.

terize disease heterogeneity and complex biological correlations, significantly enhancing predictive accuracy, stability, and generalization [106]. For instance, studies utilizing MALDI MS imaging to simultaneously capture peptide and N-glycan profiles have identified subtype-specific markers (e.g., tubulin, histone H2A, and fucosylated N-glycans). While single-mode ML models achieved accuracies of 87.49% (peptides) and 94.16% (N-glycans), multimodal data fusion further elevated accuracy to 94.25%–96.65%, representing a 4%–5.4% improvement over unimodal approaches [107]. In conclusion, the convergence of ML and glycomics offers a transformative pathway for biomarker discovery and clinical translation, particularly in the early diagnosis, stratification, and risk prediction of cancer. Multimodal integration significantly optimizes diagnostic efficacy. Future coupling with genomics, proteomics, transcriptomics, and metabolomics is expected to further enhance model precision. Notably, while these methods have demonstrated high accuracy in retrospective and cross-sectional studies, rigorous validation through prospective clinical trials remains essential for clinical implementation.

5.1.2 Biological activity screening

The biological activities of glycans are mainly mediated through two pathways: direct interactions with receptor proteins and indirect effects via active metabolites generated by gut microbial metabolism. Some high molecular weight polysaccharides (e.g., glucans) can be recognized in the gut by receptors such as EGFR, Dectin-1, and BMPRIA, absorbed into the bloodstream, and subsequently interact with cell-surface receptors to exert biological activities [107]. In contrast, macromolecular polysaccharides are often consumed as dietary components or food additives, and their physiological effects are generally mediated either directly through receptor recognition and limited absorption, or indirectly via microbial fermentation and degradation into bioactive

metabolites such as short-chain fatty acids [108]. On this basis, ML-assisted glycan bioactivity screening can be grouped into three major strategies: end-to-end bioactivity prediction, glycan-protein interaction prediction, and multi-omics-based prediction of gut microbial glycosidase profiles.

The first strategy directly predicts biological activity from glycan structures without explicitly modeling the underlying mechanism. Early studies used structural descriptors of isolated polysaccharides, such as molecular weight, monosaccharide composition, and glycosidic linkages, to predict immunomodulatory activities [109]. Although these models achieved low prediction errors, their applicability was limited by the narrow structural space of the training data. More generalizable approaches represent glycans as strings or molecular graphs. Using graph convolutional networks, glycan molecular graphs have been applied to classify immunogenicity and pathogenicity with accuracies of 0.946 and 0.919, respectively [86]. Similarly, string-based glycan representations combined with bidirectional recurrent neural networks achieved classification accuracies of 92% and 89% [93]. The relatively broad glycan coverage of such models suggests their potential for predicting the activities of larger polysaccharides. Notably, the training set for the SweetTalk model contains 19,299 glycans, with about 2000 having a degree of polymerization greater than 10 (up to 43), indicating that such methods can effectively predict polysaccharides with molecular weights of thousands of Daltons.

The second strategy screens based on whether glycans can interact with microbial glycoside hydrolases or receptor proteins. Glycan-protein interaction models usually represent glycans using molecular graphs or 3D structural features to predict binding strengths or sites. As mentioned in Section 4.4, existing models can predict binding strengths [86], but restricted by the protein types in training data, they mainly apply to viral proteins and have limited generalizability. DeepGlycanSite [91], trained on 12,023 protein-carbohydrate complex crystal structures, has broader applicability, providing an effective method for screening glycan receptor proteins and microbial glycan hydrolases in a high-throughput manner.

The third strategy integrates multi-omics data to infer the glycosidase profiles of gut microorganisms. Based on fecal metagenomic datasets from infants and adults under different dietary or prebiotic interventions, ML models have been developed to characterize polysaccharide hydrolase repertoires and distinguish the glycosidase features of different *Bifidobacterium* strains [110]. These methods provide important support for screening functional microbes capable of selectively utilizing specific polysaccharides.

5.2 Testing of polysaccharide content and crop origin

5.2.1 Rapid non-destructive quantitative analysis of polysaccharide content and composition in crops and products

Spectroscopic techniques, including near-infrared (NIR), mid-infrared (MIR), and visible-near-infrared (Vis-NIR) hyperspectral imaging, provide non-destructive tools for detecting glycan-related structural information in complex samples. By extracting informative features from spectral signals, ML enables rapid and accurate quantification of polysaccharides and their components, offering advantages of high efficiency, low cost, and non-destructive analysis. At present, these approaches have been applied mainly to the detection of polysaccharides in agricultural products and beverages.

NIR spectroscopy and Vis-NIR hyperspectral imaging combined with ML have been widely used to predict non-starch polysaccharide content in agricultural products. For example, NIR-based random forest models achieved strong predictive performance for polysaccharides in *Cistanche tubulosa*, with R^2 values above 0.86 and RPD values above 3.0 [111]. NIR combined with ML can also simultaneously predict the content of multiple polysaccharides and other substances. Portable NIR systems have also been used for simultaneous quantification of cellulose, hemicellulose, lignin, and pectin in fresh sweet potato roots [112]. In *Ganoderma lucidum*, NIR spectroscopy enabled rapid, non-destructive prediction of both total polysaccharides and moisture, while NIR hyperspectral imaging further supported polysaccharide prediction across samples from different geographical origins [113,114].

Vis-NIR hyperspectral imaging and NIR spectroscopy combined with ML have begun to be applied to predict amylose and amylopectin contents across different crop varieties. In sorghum, Vis-NIR hyperspectral imaging coupled with ML achieved highly accurate prediction of amylose and amylopectin contents [115]. Additionally, NIR-based PLSR models have successfully predicted total starch, amylose, the amylose/amylopectin ratio, neutral polysaccharides, acidic polysaccharides, and total soluble polysaccharides in yams, with all RPDs exceeding 2.5, thereby achieving simultaneous prediction for both starch and non-starch polysaccharides [116]. Beyond agricultural products, NIR and MIR spectroscopy combined with ML have also shown promise for beverage analysis. In shiitake mushroom beverages, NIR-based models achieved highly accurate prediction of lentinan content, with R^2 values above 0.985 [117]. In fresh orange juice, MIR spectroscopy combined with unsupervised and supervised ML methods enabled effective classification of soluble pectin levels, with accuracies of 89% and 85% for SVM and kNN models,

respectively [118]. Overall, the integration of spectroscopy and ML provides an efficient strategy for the rapid prediction, quantification, and classification of polysaccharides in complex food-related systems.

5.2.2 Crop origin and classification identification

This line of research utilizes fine structural differences in polysaccharides as "molecular fingerprints". By combining high-efficiency separation techniques (like LC) and highly sensitive detection methods (like MS/MS, NMR) with ML algorithms for pattern recognition, it achieves precise identification of polysaccharide origins (geographical origin, cultivar, species). Researchers extracted polysaccharides from 96 batches of *Citri Reticulatae* pericarpium across 4 varieties, utilizing HPLC to map acid hydrolysates and diffraction patterns to construct an HPLC fingerprint with 10 common peaks. An LDA model built upon this classified the varieties with > 0.954 accuracy [119]. For 205 batches of *Gastrodia elata* across 4 cultivars, 6 small molecular compounds and α1,4/1,6-glucan content were simultaneously measured. After using α-amylase-assisted Fenton's initiation to degrade the polysaccharides toward defined oligosaccharide groups and precise quantification via UHPLC and LC–MS/MS, an XGBoost classification model achieved a weighted accuracy of 0.80 on the test set, with SHAP analysis confirming α1,4/1,6-glucan as the core distinguishing indicator [120]. For 70 batches of American ginseng from 4 origins, monosaccharides and oligosaccharides were obtained via acid hydrolysis and ultrasound-assisted acid depolymerization. Monosaccharide composition/content and oligosaccharide content were acquired via HILIC-CAD, and an LDA model successfully distinguished the 4 origins with a 100% test set accuracy [121].

Collecting 140 yams across 4 species, researchers (merging two species due to small sample sizes) applied SNV processing to NIR spectra. CNN and LSTM networks extracted features from cross-sectional and powder images, and a Multimodal Transformer predicted yam species with a test set accuracy of 97.30% [116]. Using FTIR and FT-NIR spectroscopy on 262 *Gastrodia elata* samples from 3 species, a ResNet DL model achieved precise cultivar identification with 100% test set accuracy [122].

5.3 Functional modeling and smart design of polysaccharide matrices

5.3.1 Hydrogels

Polysaccharide-based hydrogels and composite gel materials exhibit excellent biocompatibility, high water retention, structural tunability, and environmental responsiveness, leading to widespread use in the food industry, wound dressings, and 3D bioprinting. Key properties such as gel density, swelling degree, mechanical strength, water vapor transmission rate, porosity, and rheological behavior are significantly influenced by raw material ratios, preparation processes, and reaction conditions. With its powerful non-linear fitting, multi-objective optimization, and non-destructive testing capabilities, ML can efficiently optimize gel formulations, control processes, predict performance, and rapidly screen candidates, offering a novel pathway for the intelligent design and precise preparation of polysaccharide-based gels.

ML can precisely predict and regulate key gel properties based on process parameters and component concentrations. Using the yeast/glucose mass ratio, gelation temperature, and reaction time as input variables, researchers built a ML model to effectively predict gel density, maximum swelling degree, and compressive strength (R^2 = 0.9287). Coupled with GA for further optimization, structurally controllable hierarchical porous hydrogels were successfully prepared [29]. Another study employed a QP-ANN model to optimize the ratio of PVA/chitosan/starch composite systems, significantly enhancing the water vapor transmission rate, gel content, swelling ratio, and porosity. This allowed the gel to maintain structural stability while efficiently absorbing wound exudate, keeping a moist healing environment, and providing good breathability, thereby meeting requirements for biological dressings and wound repair [28].

In terms of gel performance prediction and smart auxiliary testing, ML enables rapid, non-destructive, and highly efficient predictions of key indicators based on component ratios, rheological properties, and spectral data, substantially increasing material screening efficiency. Based on component ratios, an RF model using hyaluronic acid methacrylate and gelatin methacrylate concentrations predicted the system's shear rate with R^2 = 0.9739 [123]. For rheology-based printability predictions, SVR models utilizing NIR spectra and LF-NMR data as inputs successfully predicted rheological properties [124]. The NIR-based model achieved a prediction set R^2 up to 0.9796 (RPD = 7.0708), and the LF-NMR model reached R^2 = 0.9869 (RPD = 8.7220), facilitating rapid, non-destructive, indirect prediction of polysaccharide gel printability for 3D printing [125].

5.3.2 Films and coatings

Polysaccharide-based films and coatings, favored for their green safety, biodegradability, and excellent

biocompatibility, have become crucial materials in food packaging, preservation, and targeted delivery carriers. In practical applications, packaging films must simultaneously meet multiple key performance requirements, including good mechanical strength, barrier properties, and antibacterial activity. Coatings must fulfill specific needs, such as food preservation or targeted colonic release. ML provides efficient, precise solutions for formulation design, process control, and functional prediction of polysaccharide-based packaging materials.

In film preparation and optimization, ML effectively assists in optimizing raw material ratios and process parameters, significantly elevating overall film performance. Using tamarind seed starch, xanthan gum, and glycerol concentrations as inputs, a PNN model predicted the tensile strength and elongation at break of edible films with R^2 values of 0.9513 and 0.9287, respectively, providing a reliable basis for intelligent formulation optimization [27]. An ANN model utilizing starch concentration, glycerol addition, preparation temperature, and chitosan content predicted film tensile strength and elongation at break with R^2 values of 0.9955 and 0.9859 [3]. Another ANN model using ultra-high-pressure treatment pressure, essential oil concentration, and treatment time predicted the mechanical properties of potato residue edible films, achieving R^2 values of 0.9892 and 0.9866 for tensile strength and elongation at break [126]. ML can also effectively assist in optimizing and synchronously enhancing the mechanical, barrier, and antibacterial properties of polysaccharide-based packaging. Using starch concentration and orange peel essential oil addition as inputs, an ANN accurately predicted film tensile strength, water vapor transmission rate, and inhibition zone diameter (all $R^2 > 0.9999$), facilitating the optimization of a highly antibacterial composite film formulation [127].

For coating, ML can achieve precise formulation design and performance enhancement tailored to food storage requirements. Using storage time and concentrations of sodium alginate, guar gum, and pectin as inputs, SVR models predicted physicochemical indicators of strawberries (weight loss, firmness, TSS, pH, total color difference) with $R^2 > 0.97$. Subsequently, combining an overall quality score objective function with DE and PSO successfully improved strawberry quality during storage [128]. Additionally, ML coupled with spectral techniques enables the rapid prediction of packaging release behaviors, supporting targeted coating design. Utilizing Raman spectra of polysaccharides as input, a model predicted the cumulative colonic drug release rate of polysaccharide and 5-aminosalicylic acid coatings with $R^2 = 0.79$ [129].

5.3.3 Emulsions

Emulsion stability is a critical parameter affecting food shelf life, texture, and quality. Natural polysaccharides are widely used as food emulsion stabilizers due to their high safety and excellent emulsifying and stabilizing effects. As systems become increasingly complex, the efficiency of traditional evaluation methods becomes limited. Leveraging its efficient prediction and pattern mining capabilities, ML can rapidly elucidate stability mechanisms and optimize formulations, providing a highly effective new approach for regulating food emulsion stability.

ML effectively supports formulation optimization, efficient screening, and intelligent stability testing of polysaccharide-based emulsions, offering novel data-driven strategies for emulsion system design and characterization. An ANN model using pectin-rich orange by-product powder (OBF) and soy protein concentrations as inputs successfully predicted apparent viscosity, d10/d50/d90 percentiles, Zeta potential, percentage of flocculation, and creaming index ($R^2 \approx 0.997$). Following multi-constraint optimization, the optimized ratio produced an emulsion with fine particle size, no obvious phase separation, low flocculation, and excellent 24-hour stability [130].

ML also shows immense potential in the efficient screening of polysaccharide-based surfactants. Using molecular weight, glycosidic linkages, and monosaccharide composition of corn fiber gum as input features, a machine learning model predicted the TurbiScan Stability Index, Zeta potential, and particle size of emulsions, with prediction R^2 values for all three indicators $\geqslant 0.896$ [131]. Employing MOE molecular descriptors to characterize structural features alongside an RF algorithm, a QSAR classification model achieved high-throughput screening of natural emulsifiers containing polysaccharides with a 93.33% accuracy rate [132].

Furthermore, ML enables intelligent, dynamic testing of polysaccharide-based emulsion stability. Based on droplet size and count data acquired via a particle size analyzer over different storage times, an LSTM network achieved time-series prediction of cumulative droplet mass percentages ($R^2 = 0.898$), which is applicable for long-term stability assessments of gum arabic-stabilized emulsions [133]. Utilizing Confocal Laser Scanning Microscopy (CLSM) images as input, a dual-task DL model achieved statistical droplet classification (Oil-Extreme, Oil-Large, Oil-Medium, Oil-Small, Pickering-Medium, Pickering-Small) and morphological classification (Well-distributed, Flocculation, Coalescence) for Pickering emulsions, yielding mAP@0.5 scores of 0.9315 and 0.7853, respectively [134]. This offers an

efficient method for the automated quantitative analysis of emulsion microstructure and stability.

5.3.4 Cross-study comparison and generalization

Taken together, current studies demonstrate that ML can substantially improve interpolation, screening, and optimization within experimentally defined polysaccharide formulation spaces. However, there is currently insufficient evidence that one algorithm is universally superior. ANN models provide flexible nonlinear fitting but require careful regularization and validation; RF performs well on small tabular datasets but has limited extrapolation ability; SVR is suitable for small datasets but strongly depends on kernel definition and feature scaling; LSTM is advantageous only when meaningful sequential information is present; and GA, PSO, and DE can efficiently search complex surrogate response surfaces but provide no guarantee that the predicted optimum corresponds to the global optimum of the real physical system.

A further limitation is that most reported validation is interpolation-oriented. Random train/test splitting may place highly similar formulations, polysaccharide sources, spectra, or storage states in both subsets and therefore overestimate deployment performance. More rigorous evaluation should use grouped or external validation, such as leave-one-polysaccharide-source-out, leave-one-formulation-family-out, leave-one-instrument-out, or independent-laboratory testing. Model performance should also be reported together with an applicability domain or predictive uncertainty. These practices are particularly important for polysaccharides because variations in molecular-weight distribution, branching, substitution, source, purification, and processing history can produce substantial distribution shifts even when materials share the same nominal polysaccharide name.

Accordingly, the current evidence supports ML primarily as a local surrogate and screening tool rather than a universally transferable predictor. Progress toward genuinely intelligent polysaccharide material design will require models trained on structurally diverse datasets, standardized external benchmarks, explicit uncertainty assessment, and prospective experimental validation of AI-proposed formulations.

6 Challenges and Outlook

The preceding sections demonstrate that the principal bottlenecks in AI-driven glycoscience are interconnected rather than independent. Limited structural characterization constrains the quality of training data; incomplete molecular representations restrict structure–function modeling; narrow datasets limit model generalization; and insufficiently validated predictive models, in turn, restrict rational synthesis and formulation design. Therefore, future development should not be framed as isolated advances in analytical, synthetic, and computational technologies. Instead, these components should be integrated into an experimentally validated data–model–design loop.

6.1 Data scarcity is fundamentally a problem of both quantity and standardization

The first challenge is the limited availability of standardized, machine-readable glycan data. As discussed in Sections 4.2–4.4, current datasets are strongly biased toward well-characterized N- and O-linked glycans, whereas structurally heterogeneous and high-molecular-weight polysaccharides remain underrepresented. For many polysaccharides, only coarse descriptors such as molecular weight, monosaccharide composition, and branching degree are reported, and these measurements are often obtained using different analytical protocols.

Increasing data volume alone will therefore be insufficient. Experimental metadata must also be standardized so that measurements from different laboratories can be meaningfully integrated. Existing initiatives provide an important foundation. The MIRAGE initiative establishes minimum reporting requirements for glycomics experiments, including experimental procedures, analytical conditions, and data-processing information [135]. GlycoPOST further provides a FAIR repository for raw glycomics MS data and incorporates MIRAGE-compatible experimental metadata [136]. Future polysaccharide datasets should extend these principles by recording source organism, extraction and purification history, molecular-weight distribution, degree of polymerization, monosaccharide composition, linkage and branching information, chemical modifications, processing conditions, analytical platform, and batch information together with the measured properties.

High-throughput analytical methods can increase data acquisition, but analytical throughput alone does not guarantee model generalizability. CandyCrunch demonstrates that large-scale curated LC–MS/MS datasets can support high-accuracy deep-learning-based glycan identification [10]. However, its success is associated with both data scale and the structural domain represented in the training set. The same performance cannot yet be assumed for heterogeneous high-molecular-weight polysaccharides for which sequence-resolved reference data remain scarce. Thus, future data acquisition should increase both dataset size and structural coverage.

6.2 Molecular representation must bridge sequence, conformation, and macroscopic function

A second bottleneck is the mismatch between the structural information available for glycans and the information required for predictive modeling. Sequence- and graph-based representations can encode monosaccharide identity, connectivity, branching, and anomeric configuration, but they cannot fully represent conformational ensembles. Conversely, coarse descriptors such as molecular weight and monosaccharide ratios are readily measurable for polysaccharides but cannot uniquely define their molecular architectures.

Future models should therefore integrate information across multiple structural scales, including primary sequence and topology, stereochemistry, three-dimensional conformation, molecular-weight distribution, chemical modifications, and processing conditions. GlycoShape represents an important step in this direction by providing conformational information derived from molecular simulations for hundreds of glycans [81]. However, its current structural coverage remains much smaller than sequence-level glycan repositories and does not capture the full heterogeneity of high-molecular-weight polysaccharides. Consequently, 3D-informed representations should complement rather than replace experimentally derived structural descriptors.

For polysaccharide matrices, molecular descriptors and formulation variables should also be explicitly distinguished. A model based only on polymer concentration may accurately describe one experimental system but cannot distinguish concentration effects from differences in molecular weight, substitution, branching, or source. Integrating molecular structure with processing and formulation variables would therefore shift ML from empirical formulation fitting toward more transferable structure–process–property relationships.

6.3 Generalization and uncertainty should become primary evaluation criteria

A third major challenge is model evaluation. As shown in Section 5.3, high R^2, accuracy, or AUC values are frequently reported, but these metrics are commonly derived from random partitions of datasets generated within a single laboratory, material source, or experimental design. Such validation primarily evaluates interpolation and provides limited evidence that a model will perform reliably for previously unseen glycans, polysaccharide sources, or processing conditions.

Future benchmarks should therefore deliberately evaluate models under distribution shifts. GlycanML provides a useful example by constructing standardized benchmarks for multiple glycan prediction tasks and using motif-based dataset splitting to evaluate models on glycans containing structural patterns distinct from those encountered during training [123]. Similar principles could be extended to polysaccharide research through leave-one-source-out, leave-one-structural-family-out, leave-one-batch-out, leave-one-instrument-out, and independent-laboratory validation. Such evaluation would provide a more realistic estimate of model performance when applied to newly characterized polysaccharides.

Predictive uncertainty should also accompany point predictions, particularly when ML models are coupled with optimization algorithms. In inverse design, an optimizer may preferentially search regions where a surrogate model predicts extreme performance even though these regions are poorly represented in the training data. This can lead to apparently optimal solutions that reflect model extrapolation errors rather than genuinely superior experimental conditions. Uncertainty-calibrated reaction prediction has demonstrated that model confidence can provide useful information on prediction reliability [137]. Extending uncertainty estimation and applicability-domain analysis to glycan and polysaccharide models would therefore improve the reliability of candidate selection and experimental decision-making.

6.4 Connecting prediction with synthesis and experimental validation

The final challenge is to convert predictive capability into experimentally actionable glycan design. This connection is becoming increasingly feasible because advances in glycan synthesis have expanded the experimentally accessible structural space. Automated glycan assembly has enabled the synthesis of a sequence-defined 100-mer polysaccharide [138], while automated solution-phase multiplicative synthesis has subsequently extended chemically synthesized glycan chains to 1080 monosaccharide units [139]. These advances indicate that increasingly large and structurally defined glycans can be prepared experimentally, providing an important basis for establishing systematic sequence–structure–property datasets.

At the computational level, data-efficient approaches may be more practical than simply increasing model complexity. Pesciullesi et al. demonstrated that transfer learning from a general chemical reaction model to a smaller carbohydrate-specific reaction dataset substantially improved the prediction of regio- and stereoselective carbohydrate transformations and experimentally validated the resulting model during oligosaccharide synthesis [140]. This study provides direct evidence that information learned from larger chemical reaction datasets can be transferred to data-scarce carbohydrate

chemistry. However, the effectiveness of transfer learning will depend on the similarity between the source and target chemical domains and should therefore be evaluated rather than assumed.

The long-term objective should consequently be a closed-loop workflow rather than independent advances in analytical characterization, synthesis, and AI. High-quality structural and functional measurements should first be deposited using standardized representations and metadata. These data can then support predictive or generative models that propose candidate glycan structures, reaction conditions, or formulations. Automated or semi-automated synthesis and experimental characterization can subsequently test these candidates, while both successful and unsuccessful experiments are returned to the dataset for model updating. This framework directly connects the analytical strategies discussed in Sections 4.2–4.4, the synthesis and reaction-prediction approaches described in Section 4.1.2, the inverse-design concepts discussed in Section 4.5, and the formulation optimization strategies summarized in Section 5.3.

Accordingly, the immediate priority for AI-driven glycoscience is not necessarily the development of increasingly complex algorithms. More critical needs are structurally diverse and standardized datasets, evaluation protocols that explicitly test out-of-distribution generalization, uncertainty-aware predictions, and prospective experimental validation. Addressing these issues would allow the field to progress from high-accuracy retrospective fitting toward reproducible, transferable, and experimentally actionable glycan design.

7 Conclusions and Future Researches

In summary, machine learning is reshaping glycan research by providing powerful tools for handling high-dimensional, nonlinear, and heterogeneous data that are difficult to address using conventional experimental and statistical approaches. Across the full research pipeline, from extraction and synthesis to sequence elucidation, 3D structures prediction, interaction analysis, and property-guided molecular design, ML has demonstrated substantial potential to accelerate discovery and improve precision. In particular, the integration of ML with NMR, MS, nanopore sensing, spectroscopy, imaging, and glycomics has greatly expanded the ability to characterize complex glycans and uncover their functional relevance in food science, biomedicine, and biomaterials.

Despite the rapid progress of AI-driven glycoscience, several challenges remain, including the limited availability of large-scale standardized datasets, the structural bias of existing data toward well-studied glycan classes, and the continued reliance on coarse-grained descriptors that fail to fully capture sequence and conformational complexity. Future advances will rely on the integration of high-throughput analytical technologies, improved glycan synthesis, and next-generation AI approaches such as transfer learning, multimodal learning, and physics-informed modeling, collectively driving glycoscience from descriptive prediction toward mechanism-based understanding and rational design.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments This work was supported by the National Key Research and Development Program of China (2024YFF1106000) and the Guangdong Basic and Applied Basic Research Foundation (2023B1515040014). The authors used ChatGPT (OpenAI, GPT-5.6, online version) solely as a language-editing tool to improve grammar, clarity, and readability. The AI tool was not used for data analysis, data generation, image processing, or scientific content creation. All outputs were carefully reviewed and verified by the authors, who take full responsibility for the final content.

Competing interests The authors declare no conflict of interest.

References

1. Liu Y N, Huang Y J, Zhu R Y, et al. Structural elucidation approaches in carbohydrates: a comprehensive review on techniques and future trends. Food Chemistry, 2023, 400: 134118
2. Yao H Y Y, Wang J Q, Yin J Y, et al. A review of NMR analysis in polysaccharide structure and conformation: progress, challenge and perspective. Food Research International, 2021, 143: 110290
3. Tan S X, Ong H C, Andriyana A, et al. Characterization and parametric study on mechanical properties enhancement in biodegradable chitosan-reinforced starch-based bioplastic film. Polymers, 2022, 14(2): 278
4. Zhang J X, Xu Z Y, Zhong C, et al. Artificial intelligence reshaping non-starch polysaccharides research: a comprehensive review of intelligent integration from extraction optimization to functional design. Comprehensive Reviews in Food Science and Food Safety, 2026, 25(1): e70379
5. Lin N, Ren Y Q, Sun D W. Advancements in novel polysaccharide analysis: emerging spectral technology approaches. Trends in Food Science & Technology, 2025, 162: 105037
6. Varki A. Biological roles of glycans. Glycobiology, 2017, 27(1): 3–49
7. Wang J Q, Zhao J, Nie S P, et al. MALDI mass spectrometry in food carbohydrates analysis: a review of recent researches. Food Chemistry, 2023, 399: 133968
8. Abramson J, Adler J, Dunger J, et al. Accurate structure

prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630(8016): 493–500
9. Bagdonaite I, Malaker S A, Polasky D A, et al. Glycoproteomics. Nature Reviews Methods Primers, 2022, 2(1): 48
10. Urban J, Jin C S, Thomsson K A, et al. Predicting glycan structure from tandem mass spectrometry via deep learning. Nature Methods, 2024, 21(7): 1206–1215
11. Yao G D, Ke W J, Xia B Q, et al. Nanopore-based glycan sequencing: state of the art and future prospects. Chemical Science, 2024, 15(17): 6229–6243
12. Djalali S, Yadav N, Delbianco M. Towards glycan foldamers and programmable assemblies. Nature Reviews Materials, 2024, 9(3): 190–201
13. Ogawa Y, Putaux J L, Nishiyama Y. Crystallography of polysaccharides: current state and challenges. Current Opinion in Chemical Biology, 2022, 70: 102183
14. Angulo J, Ardá A, Bertuzzi S, et al. NMR investigations of glycan conformation, dynamics, and interactions. Progress in Nuclear Magnetic Resonance Spectroscopy, 2024, 144–145: 97–152
15. Huang J H, Tao H, Chen J K, et al. Structure-guided discovery of protein and glycan components in native mastigonemes. Cell, 2024, 187(7): 1733–1744. e12
16. Tsai Y X, Chang N E, Reuter K, et al. Rapid simulation of glycoprotein structures by grafting and steric exclusion of glycan conformer libraries. Cell, 2024, 187(5): 1296–1311. e26
17. Wang T T, Huang W Z, Xu K, et al. CryoSeek II: cryo-EM analysis of glycofibrils from freshwater reveals well-structured glycans coating linear tetrapeptide repeats. Proceedings of the National Academy of Sciences of the United States of America, 2024, 122(1): e2423943122
18. Kirschner K N, Yongye A B, Tschampel S M, et al. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. Journal of Computational Chemistry, 2008, 29(4): 622–655
19. AlYammahi J, Darwish A S, Lemaoui T, et al. Parametric analysis and machine learning for enhanced recovery of high-value sugar from date fruits using supercritical CO_2 with co-solvents. Journal of CO_2 Utilization, 2023, 72: 102511
20. Chen H Y, Wang B, Li J P, et al. Comparative study on the extraction efficiency, characterization, and bioactivities of *Bletilla striata* polysaccharides using response surface methodology (RSM) and genetic algorithm-artificial neural network (GA-ANN). International Journal of Biological Macromolecules, 2023, 226: 982–995
21. Imran M, Weng C H, Wolde G S, et al. Ultrasound-assisted deep eutectic solvent-based extraction of polysaccharides from okra: optimization by response surface methodology and artificial neural network modeling. Ultrasonics Sonochemistry, 2026, 124: 107715
22. Li Z Y, Li Q, Hu C G, et al. The machine learning assisted optimization of ultrasound-synergistic enzymatic extraction and comprehensive evaluation of polysaccharides from *Saposhnikovia divaricata*. Ultrasonics Sonochemistry, 2026, 124: 107706
23. Liu F, Liu W H, Tian S G. Artificial neural network optimization of *Althaea rosea* seeds polysaccharides and its antioxidant activity. International Journal of Biological Macromolecules, 2014, 70: 100–107
24. Qiu J Y, Kou R B, Zeng S Y, et al. Ionic liquid-mediated ultrasonic-assisted extraction of *Curcuma phaeocaulis* polysaccharides: process optimization, structural characterization, and extraction mechanisms. LWT, 2026, 242: 119123
25. Wang D Y, Zhang M, Law C L, et al. Natural deep eutectic solvents for the extraction of lentinan from shiitake mushroom: COSMO-RS screening and ANN-GA optimizing conditions. Food Chemistry, 2024, 430: 136990
26. Ramos-Frutos J, Cruz-Salgado, Ramos-Soto, et al. Improving the response surface methodology optimization with metaheuristics: A practical approach . Expert Systems with Applications, 2026, 297: 129195
27. Chandra Mohan C, Rakhavan K R, Sudharsan K, et al. Design and characterization of spice fused tamarind starch edible packaging films. LWT - Food Science and Technology, 2016, 68: 642–652
28. Joorabloo A, Khorasani M T, Adeli H, et al. Using artificial neural network for design and development of PVA/chitosan/starch/heparinized nZnO hydrogels for enhanced wound healing. Journal of Industrial and Engineering Chemistry, 2022, 108: 88–100
29. Yang S L, Tian X K, Zhang Q S, et al. Microorganism inspired hydrogels: optimization by response surface methodology and genetic algorithm based on artificial neural network. European Polymer Journal, 2023, 201: 112497
30. Sun D W, Pu H B, Yu J X. Applications of hyperspectral imaging technology in the food industry. Nature Reviews Electrical Engineering, 2024, 1(4): 251–263
31. Biau G, Scornet E. A random forest guided tour. TEST, 2016, 25(2): 197–227
32. Strobl C, Boulesteix A L, Zeileis A, et al. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics, 2007, 8(1): 25
33. Smola A J, Schölkopf B. A tutorial on support vector regression. Statistics and Computing, 2004, 14(3): 199–222
34. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, et al. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing, 2020, 408: 189–215
35. Tharwat A, Gaber T, Ibrahim A, et al. Linear discriminant analysis: a detailed tutorial. AI Communications, 2017, 30(2): 169–190
36. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 2021, 54(3): 1937–1967
37. Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining. San Francisco, USA, 2016, 785–794
38. Mehmood T, Ahmed B. The diversity in the applications of partial least squares: an overview. Journal of Chemometrics, 2016, 30(1): 4–17
39. Cunningham P, Delany S J. k-nearest Neighbour classifiers - a tutorial. ACM Computing Surveys, 2022, 54(6): 128
40. Gao J L, Wang S J, Erdogdu F, et al. Machine learning-driven innovations in food processing: a systematic review of applications, challenges, and future developments. Trends in Food Science & Technology, 2025, 164: 105204
41. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444
42. Specht D F. Probabilistic neural networks. Neural Networks, 1990, 3(1): 109–118
43. Greff K, Srivastava R K, Koutník J, et al. LSTM: a search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2222–2232
44. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780
45. Elman J L. Finding structure in time. Cognitive Science, 1990, 14(2): 179–211
46. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: nonlinear Phenomena, 2020, 404: 132306
47. Zhang S, Tong H H, Xu J J, et al. Graph convolutional networks: a comprehensive review. Computational Social Networks, 2019, 6(1): 11
48. Tay Y, Dehghani M, Bahri D, et al. Efficient transformers: a survey. ACM Computing Surveys, 2023, 55(6): 109
49. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017, 6000–6010
50. Yu F, Koltun V, Funkhouser T. Dilated residual networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017, 636–644
51. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324
52. Zhao J, Yu Y, Lu S H, et al. Convolutional neural networks in the realm of food quality and safety evaluation: current achievements and future prospects. Trends in Food Science & Technology, 2025, 163: 105162
53. Whitley D. A genetic algorithm tutorial. Statistics and Computing, 1994, 4(2): 65–85
54. Wolpert D H, Macready W G. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 67–82
55. Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182–197
56. Clerc M, Kennedy J. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. IEEE Transactions on Evolutionary Computation, 2002, 6(1): 58–73
57. Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of ICNN'95−International Conference on Neural Networks. Perth, Australia, 1995, 1942–1948
58. Das S, Suganthan P N. Differential evolution: a survey of the state-of-the-art. IEEE Transactions on Evolutionary Computation, 2011, 15(1): 4–31
59. Zhong C T, Li G, Meng Z, et al. Starfish optimization algorithm (SFOA): a bio-inspired metaheuristic algorithm for global optimization compared with 100 optimizers. Neural Computing and Applications, 2025, 37(5): 3641–3683
60. Mirjalili S, Lewis A. The whale optimization algorithm. Advances in Engineering Software, 2016, 95: 51–67
61. Yan Z P, Zhang J Z, Zeng J, et al. Nature-inspired approach: an enhanced whale optimization algorithm for global optimization. Mathematics and Computers in Simulation, 2021, 185: 17–46
62. Xue J K, Shen B. A novel swarm intelligence optimization approach: sparrow search algorithm. Systems Science & Control Engineering, 2020, 8(1): 22–34
63. Ma J M, Yao J L, Ren X Y, et al. Machine learning-assisted data-driven optimization and understanding of the multiple stage process for extraction of polysaccharides and secondary metabolites from natural products. Green Chemistry, 2023, 25(8): 3057–3068
64. Yao X W, Ma A T, Wang G J, et al. Optimization of ultrasonic extraction and kinetic modeling of *jieduquyuziyin* prescription polysaccharides via explainable machine learning. Ultrasonics Sonochemistry, 2025, 122: 107635
65. Tian Y S, Yang X, Chen N H, et al. Data-driven interpretable analysis for polysaccharide yield prediction. Environmental Science and Ecotechnology, 2024, 19: 100321
66. Amenaghawon A N, Igemhokhai S, Eshiemogie S A, et al. Data-driven intelligent modeling, optimization, and global sensitivity analysis of a xanthan gum biosynthesis process. Heliyon, 2024, 10(3): e25432
67. Karmakar B, Saha S P, Chakraborty R, et al. Optimization of starch extraction from *Amorphophallus paeoniifolius* corms using response surface methodology (RSM) and artificial neural network (ANN) for improving yield with tenable chemical attributes. International Journal of Biological Macromolecules, 2023, 237: 124183
68. Sahu N, Mahanty B, Haldar D. Challenges and opportunities in bioprocessing of gellan gum: a review. International Journal of Biological Macromolecules, 2024, 276: 133912
69. Xue H K, Xu J Q, Zhang J L, et al. Modeling, optimization, purification, and characterization of polysaccharides from *Lilium lancifolium* Thunb. LWT, 2022, 162: 113491
70. Deng G F, Li P J, Liang H, et al. Extraction of polysaccharides from *Camellia oleifera* leaves by dual enzymes combined with deep eutectic solvents screened by ANN and COSMO-RS. International Journal of Biological Macromolecules, 2025, 305: 141131
71. Astray G, Gullón B, Labidi J, et al. Comparison between developed models using response surface methodology (RSM) and artificial neural networks (ANNs) with the purpose to

optimize oligosaccharide mixtures production from sugar beet pulp. Industrial Crops and Products, 2016, 92: 290–299
72. Sabater C, Sabater V, Olano A, et al. Ultrasound-assisted extraction of pectin from artichoke by-products. An artificial neural network approach to pectin characterisation. Food Hydrocolloids, 2020, 98: 105238
73. Faurschou N V, Friis V, Raghavan P, et al. Anomeric selectivity of glycosylations through a machine learning lens. Journal of the American Chemical Society, 2025, 147(40): 36197–36209
74. Wang Y H, Han S X, Wang Y, et al. Artificial intelligence technology assists enzyme prediction and rational design. Journal of Agricultural and Food Chemistry, 2025, 73(12): 7065–7073
75. Pereira F. 1D ^{13}C-NMR data as molecular descriptors in spectra—structure relationship analysis of oligosaccharides. Molecules, 2012, 17(4): 3818–3833
76. Gao X M, Xiao Y Q, Li W Z, et al. Nanopore analysis of malto-oligosaccharides from amylose, amylopectin, and cyclodextrin hydrolysates. Nano Letters, 2026, 26(1): 74–82
77. Xia K, Hagan J T, Fu L, et al. Synthetic heparan sulfate standards and machine learning facilitate the development of solid-state nanopore analysis. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(11): e2022806118
78. Zong Y, Wang Y X, Qiu X P, et al. Deep learning prediction of glycopeptide tandem mass spectra powers glycoproteomics. Nature Machine Intelligence, 2024, 6(8): 950–961
79. Yang Y, Fang Q. Prediction of glycopeptide fragment mass spectra by deep learning. Nature Communications, 2024, 15(1): 2448
80. Sun W P, Zhang Q Q, Zhang X Y, et al. Glycopeptide database search and de novo sequencing with PEAKS GlycanFinder enable highly sensitive glycoproteomics. Nature Communications, 2023, 14(1): 4046
81. Ives C M, Singh O, D'Andrea S, et al. Restoring protein glycosylation with GlycoShape. Nature Methods, 2024, 21(11): 2117–2127
82. Fadda E. Molecular simulations of complex carbohydrates and glycoconjugates. Current Opinion in Chemical Biology, 2022, 69: 102175
83. Woods R J. Predicting the structures of glycans, glycoproteins, and their complexes. Chemical Reviews, 2018, 118(17): 8005–8024
84. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nature Machine Intelligence, 2021, 3(12): 1023–1032
85. Thomès L, Joeres R, Akdeniz Z, et al. GlyContact analyzes glycan 3D structures at scale. Nature Communications, 2025, 16(1): 11136
86. Burkholz R, Quackenbush J, Bojar D. Using graph convolutional neural networks to learn a representation for glycans. Cell Reports, 2021, 35(11): 109251
87. Carpenter E J, Seth S, Yue N, et al. GlyNet: a multi-task neural network for predicting protein–glycan interactions. Chemical Science, 2022, 13(22): 6669–6686
88. Carpenter E J, Peng C H, Haregu S, et al. Atom-level machine learning of protein-glycan interactions and cross-chiral recognition in glycobiology. Science Advances, 2025, 11(49): eadx6373
89. Banno M, Komiyama Y, Cao W, et al. Development of a sugar-binding residue prediction system from protein sequences using support vector machine. Computational Biology and Chemistry, 2017, 66: 36–43
90. Taherzadeh G, Zhou Y Q, Liew A W C, et al. Sequence-based prediction of protein–carbohydrate binding sites using support vector machines. Journal of Chemical Information and Modeling, 2016, 56(10): 2115–2122
91. He X H, Zhao L F, Tian Y P, et al. Highly accurate carbohydrate-binding site prediction with DeepGlycanSite. Nature Communications, 2024, 15(1): 5163
92. Chu Q Y, He X H, Tan X Y, et al. GlycanInsight: an open platform for carbohydrate-binding pocket prediction and characterization. Chemical Science, 2025, 16(23): 10264–10272
93. Bojar D, Powers R K, Camacho D M, et al. Deep-learning resources for studying glycan-mediated host-microbe interactions. Cell Host & Microbe, 2021, 29(1): 132–144. e3
94. Mufti Z S, Asim M, Shflot A S, et al. Data-driven regression analysis of amylose using Sombor molecular descriptors. Scientific Reports, 2025, 15(1): 44294
95. Liao H G, Hu S, Yang H, et al. Data-driven de novo design of super-adhesive hydrogels. Nature, 2025, 644(8075): 89–95
96. Reily C, Stewart T J, Renfrow M B, et al. Glycosylation in health and disease. Nature Reviews Nephrology, 2019, 15(6): 346–366
97. Wang Y, Feng Y T, Li J Y, et al. Immunoglobulin G N-glycan signatures as potential diagnostic and predictive biomarkers for non-small-cell lung cancer. International Journal of Biological Macromolecules, 2025, 320: 146089
98. Zhang G H, Huang X D, Gong Y L, et al. Fingerprint profiling of glycans on extracellular vesicles via lectin-induced aggregation strategy for precise cancer diagnostics. Journal of the American Chemical Society, 2024, 146(42): 29053–29063
99. Wang Y L, Lai Y, Wang J R, et al. Serum IgG N-glycans act as serum biomarkers for differentiation of cold and heat pattern in rheumatoid arthritis. Chinese Medicine, 2025, 20(1): 184
100. Fu B, Chen J R, Liu X J, et al. Large-scale serum N-glycomics tracks N-glycosylation dynamics in hepatocellular carcinoma progression and enables early diagnosis. Nature Communications, 2026, 17(1): 1885
101. Vathy-Fogarassy A, Gombas V, Torok R, et al. Improved analytical workflow towards machine learning supported N-glycomics-based biomarker discovery. Talanta, 2025, 295: 128389
102. Huang C J, Fang M, Feng H J, et al. N-glycan fingerprint predicts alpha-fetoprotein negative hepatocellular carcinoma: a large-scale multicenter study. International Journal of Cancer, 2021, 149(3): 717–727
103. Wu Z Y, Guo Z, Zheng Y L, et al. IgG N-glycosylation

cardiovascular age tracks cardiovascular risk beyond calendar age. Engineering, 2023, 26: 99–107
104. Lin Y W, Li N, Zhang H Y H, et al. Hierarchical channeled graphitized nanoarchitecture as a diagnostic platform for maternal fever warning. Advanced Science, 2026, 13(26): e21251
105. Mészáros B, Járvás G, Kun R, et al. Machine learning based analysis of human serum N-glycome alterations to follow up lung tumor surgery. Cancers, 2020, 12(12): 3700
106. Wang X, Nguyen J, Nader K, et al. Discovery of predictive biomarkers for cancer therapy through computational approaches. Nature Reviews Clinical Oncology, 2026, 23(3): 180–200
107. Bollwein C, Patterson N H, Ly A, et al. Classification of small blue round cell tumors by integrating peptide and N-glycan mass imaging spectrometric profiles. Analytica Chimica Acta, 2025, 1376: 344614
108. Liao W F, Cao D X, Wang Y, et al. 1,3-and 1,4-linked polysaccharides uptake in intestinal cells relies on clathrin/dynamin 1/Rab5-dependent endocytosis. Nature Communications, 2026, 17(1): 1831
109. Lu J, Yang Y J, Hong E K, et al. Analyzing the structure-activity relationship of raspberry polysaccharides using interpretable artificial neural network model. International Journal of Biological Macromolecules, 2024, 264: 130354
110. Sabater C, Ruiz L, Margolles A. A machine learning approach to study glycosidase activities from *Bifidobacterium*. Microorganisms, 2021, 9(5): 1034
111. Wang Y, Tian Z P, Xie J J, et al. Rapid determination of polysaccharides in *Cistanche Tubulosa* using near-infrared spectroscopy combined with machine learning. Journal of AOAC INTERNATIONAL, 2023, 106(4): 1118–1125
112. He H J, Zhang C, Xing L Z, et al. Machine learning-driven pocket-sized NIR analysis for non-invasive quantification of cellulose, hemicellulose, lignin and pectin in sweet potato roots. LWT, 2025, 226: 117973
113. Ni H F, Fu W L, Wei J, et al. Non-destructive detection of polysaccharides and moisture in *Ganoderma lucidum* using near-infrared spectroscopy and machine learning algorithm. LWT, 2023, 184: 115001
114. Ran J, Xu H, Wang Z L, et al. Non-destructive analysis of *Ganoderma lucidum* composition using hyperspectral imaging and machine learning. Frontiers in Chemistry, 2025, 13: 1534216
115. Huang H P, Fei X, Hu X J, et al. Analysis of the spectral and textural features of hyperspectral images for the nondestructive prediction of amylopectin and amylose contents of sorghum. Journal of Food Composition and Analysis, 2023, 115: 105018
116. Fan X Q, Wei M H, Yang L, et al. Rapid identification and prediction of the botanical origins and main components of Chinese yam based on a multimodal data fusion of near-infrared spectra and visualization image information. Food Chemistry, 2026, 511: 148819
117. Wang D Y, Zhang M, Adhikari B, et al. Determination of polysaccharide content in shiitake mushroom beverage by NIR spectroscopy combined with machine learning: a comparative analysis. Journal of Food Composition and Analysis, 2023, 122: 105460
118. Bizzani M, William Menezes Flores D, Alberto Colnago L, et al. Monitoring of soluble pectin content in orange juice by means of MIR and TD-NMR spectroscopy combined with machine learning. Food Chemistry, 2020, 332: 127383
119. Zhong M Y, Li M N, Zou W S, et al. Differentiation of Citri Reticulatae Pericarpium varieties via HPLC fingerprinting of polysaccharides combined with machine learning. Food Chemistry, 2025, 473: 143053
120. Kong X Q, Chen S H, Shen L L, et al. Polysaccharide-driven quality control of *Gastrodia elata* Blume: a multidimensional strategy integrating LC–MS/MS-based quantitative analysis, machine learning, and anti-neuroinflammatory bioactivity. Journal of Pharmaceutical and Biomedical Analysis, 2026, 275: 117457
121. Hu S M, Guo H Y, Yuan Y, et al. Machine learning combined with fusion of multiple sources of polysaccharide data based on hydrophilic interaction chromatography-charged aerosol detection for geographical origin discrimination of American ginseng (*Panax quinquefolius*). Journal of Food Composition and Analysis, 2025, 144: 107729
122. Li G Y, Li J Q, Liu H G, et al. Rapid and accurate identification of *Gastrodia elata* Blume species based on FTIR and NIR spectroscopy combined with chemometric methods. Talanta, 2025, 281: 126910
123. Deng B C, Chen S B, Lasaosa F L, et al. Predicting rheological properties of HAMA/GelMA hybrid hydrogels via machine learning. Journal of the Mechanical Behavior of Biomedical Materials, 2025, 168: 107005
124. Lu Y X, Rai R, Nitin N. Image-based assessment and machine learning-enabled prediction of printability of polysaccharides-based food ink for 3D printing. Food Research International, 2023, 173: 113384
125. Tang T T, Zhang M, Adhikari B, et al. Indirect prediction of the 3D printability of polysaccharide gels using multiple machine learning (ML) models. International Journal of Biological Macromolecules, 2024, 280: 135769
126. Akdemir Evrendilek G, Bulut N, Uzuner S. Assessing edible composite film polymer from potato industry effluent under high hydrostatic pressure and its antimicrobial properties. CLEAN - Soil Air Water, 2022, 50(11): 2100406
127. Waday Y A, Aklilu E G. ANN and RSM modeling for the synthesis of avocado seed starch combined orange peel extract antimicrobial packaging film. International Journal of Polymer Science, 2023, 2023(1): 8877598
128. Niam S, Ahmad I, Rayhan M A, et al. Machine learning-based optimization of alginate, guar gum, and pectin-based edible coatings for extended strawberry shelf life. LWT, 2025, 233: 118548
129. Abdalla Y, McCoubrey L E, Ferraro F, et al. Machine learning of Raman spectra predicts drug release from polysaccharide

coatings for targeted colonic delivery. Journal of Controlled Release, 2024, 374: 103–111

130. Umaña M, Llull L, Bon J, et al. Artificial neural networks to optimize oil-in-water emulsion stability with orange by-products. Foods, 2022, 11(23): 3750
131. Yang Y Q, Liu H H, Zhu Q M, et al. Advancing structure–emulsifying property relationships of corn fiber gum–soy protein isolate conjugates through machine learning. Food Hydrocolloids, 2026, 178: 112664
132. Liu M Q, Zhou Z X, Zhao H F, et al. Identification of natural food-derived emulsifiers using QSAR and machine learning: application in dairy emulsions. Food Chemistry, 2025, 491: 145260
133. Saddiqi H A, Ullah A, Javed Z, et al. Predictive modeling of emulsion stability and drop characteristics using machine learning: a study on surfactant influence and time dynamics. Food and Bioproducts Processing, 2025, 149: 190–198
134. Huang Z Y, Ni Y, Yu Q, et al. Deep learning in food science: an insight in evaluating Pickering emulsion properties by droplets classification and quantification via object detection algorithm. Advances in Colloid and Interface Science, 2022, 304: 102663
135. York W S, Agravat S, Aoki-Kinoshita K F, et al. MIRAGE: the minimum information required for a glycomics experiment. Glycobiology, 2014, 24(5): 402–406
136. Watanabe Y, Aoki-Kinoshita K F, Ishihama Y, et al. GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. Nucleic Acids Research, 2021, 49(D1): D1523–D1528
137. Schwaller P, Laino T, Gaudin T, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Central Science, 2019, 5(9): 1572–1583
138. Joseph A A, Pardo-Vargas A, Seeberger P H. Total synthesis of polysaccharides by automated glycan assembly. Journal of the American Chemical Society, 2020, 142(19): 8561–8564
139. Yao W L, Xiong D C, Yang Y, et al. Automated solution-phase multiplicative synthesis of complex glycans up to a 1,080-mer. Nature Synthesis, 2022, 1(11): 854–863
140. Pesciullesi G, Schwaller P, Laino T, et al. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. Nature Communications, 2020, 11(1): 4874